University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 130

Long-Term Commitment and Cooperation

Frédéric Schneider and Roberto A. Weber

September 2013

Long-Term Commitment and Cooperation

Frédéric Schneider* Roberto A. Weber*

September 30, 2013

Abstract

We study how the willingness to enter long-term bilateral relationships affects cooperation even when parties have little information about each other, *ex ante*, and cooperation is otherwise unenforceable. We experimentally investigate a finitely-repeated Prisoner's Dilemma, allowing players to endogenously select interaction durations. Consistent with prior research, longer interactions facilitate cooperation. However, many individuals avoid long-term commitment, with uncooperative types less likely to commit than conditional cooperators. Endogenously chosen long-term commitment yields higher cooperation rates (98% in one condition) than exogenously imposed commitment. Thus, the willingness to enter into long-term relationships provides a means for fostering—and screening for—efficient cooperation.

∗. Department of Economics, University of Zürich, Blümlisalpstrasse 10, 8006 Zürich. FS (corresponding author): frederic.schneider@econ.uzh.ch; RW: roberto.weber@econ.uzh.ch.

1. Introduction

The problem of how to obtain voluntary cooperation—i.e., when cooperation is otherwise difficult to achieve, perhaps because relevant behaviors are not contractible or directly observable—is central to the social sciences. Finding simple and generalizable ways to facilitate cooperation among interacting individuals is of primary importance in social policy and in organizational and market design.

One natural way to facilitate cooperation is through long-term interaction between individuals, as when parties commit to each other contractually for a long period of time (Friedman 1971; Kreps et al. 1982). Indeed, laboratory experiments that exogenously vary the interaction horizon of players in prisoner's dilemma, gift-exchange, and public good games (Andreoni and Miller 1993; Gächter and Falk 2002; Dal Bó 2005) show that longer (expected) interaction between parties facilitates cooperation.[1]

In this paper, we study the *endogenous* development of long-term interaction and its effects on cooperation. While the benefits of long-term interaction for facilitating cooperation are well established, an important open question is the extent to which individuals will voluntarily agree to lengthy contracts that commit them to interacting with the same counterparts for long periods of time. That is, are people willing to voluntarily commit to long-term interactions? The benefits of doing so might be less salient than the potential risk of committing to interacting with an uncooperative counterpart, and the resulting possibility of exploitation.

Moreover, if people do voluntarily commit to long-term interaction, what effect does such endogenous determination of interaction duration have on voluntary cooperation, relative to when it is exogenously imposed? For example, do different types of people differentially select long-term interaction, resulting in different samples and degrees of cooperative behavior, relative to when everyone is in long-term interaction? Or, even if there is no differential selection, does the act of having voluntarily chosen long-term commitment make a pair more likely to cooperate? These questions are important because

1. Camera and Casari (2009) show that private reputations and punishment threats are key factors to uphold cooperation in repeated prisoner's dilemmas.

the answers will indicate whether people can use willingness to commit as a way to screen their interaction partners for cooperativeness.

To answer these questions, we use a laboratory experiment with a finitely-repeated Prisoner's Dilemma (PD) game to study exogenously imposed and endogenously selected long-term interaction. To abstract from other features of interaction that may affect the choice of long-term interaction and resulting cooperation, we construct a highly stylized and simple environment in which players are provided with no information on opponents prior to selecting whether to commit to one selected at random. Thus, our experiment differs substantially from other related research that considers "partner selection" or "break-ups" of existing relationships in Prisoner's Dilemma games (Hauk and Nagel 2001; Page, Putterman, and Unel 2005; Fujiwara-Greve and Okuno-Fujiwara 2009; Jackson and Watts 2010) and in which players decide whether to (continue to) play with other players after observing how cooperative they have been in the past.[2]

We begin with experimental conditions in which we exogenously vary the finite time horizon confronting interacting players. Consistent with prior research, we find that longer interaction horizons yield significantly higher cooperation rates.

We then study, as our focus, conditions in which individuals choose for how long they want to interact with a randomly-determined opponent. That is, in conditions with endogenous interaction duration, subjects do not select with whom to interact, but rather simply for how long they wish to interact with someone, selected at random. We vary the length of contracts available, with in some cases players having the option to commit to play with the same counterpart for the entire experiment. Importantly, while the interaction horizon varies, the stage game is always the same in our experiment.

We also elicit, separately, a measure of each subject's intrinsic cooperativeness and belief about the prevalence of cooperative types in the population. These measures allow us to independently identify a subject's social "type," as a way to investigate how differences in cooperativeness and beliefs affect the endogenous selection of interaction durations.

2. For related theoretical work, see Ghosh and Ray (1996) and Rob and Yang (2010).

Our experiment is motivated primarily by three questions. First, when people have the option to commit to long-term interaction with the same counterpart, do they voluntarily agree to such an option? Given the benefits of exogenously imposed long interaction horizons for cooperation, are such benefits realized when people decide for themselves for how long to interact? Or, does a "fear of commitment"—for instance, an aversion to the possibility of being paired with someone who always defects—lead people to select short-term contracts that are less effective for facilitating cooperation?

Second, we ask *who* opts for long-term commitment—is it those who are the most or least inclined to cooperate? People who like mutual cooperation may prefer long-term commitment, because they can use repeated game incentives to discipline potential defectors and induce mutual cooperation. On the other hand, selfish people may prefer a one-shot environment if they think there are enough "suckers" on whom they can defect with impunity. So, an environment in which people can freely choose whether or not to commit to long-term interaction may lead to self-selection according to people's social types.

Finally, a third central question deals with the effects of endogenously-chosen interaction durations on cooperation. Do people who voluntarily agree to interact for longer periods of time cooperate more than those for whom similarly long interaction horizons are exogenously imposed? There are at least two reasons why this might be the case. First, selection by cooperative types into long-term contracts may make these contracts better for fostering cooperation than when everyone is forced into such contracts. Additionally, the willingness to enter into such contracts may convey additional information that facilitates cooperation. That is, players' willingness to commit to long-term interaction may provide a signal of the action that they intend to play in a manner that further facilitates subsequent cooperation. So even if the type distribution is kept constant, the "institution" of endogenous commitment could foster cooperation. We explore both possibilities in our data.

Previewing our results, our data corroborate earlier findings that longer interaction horizons increase average cooperation rates. However, our experiment also reveals that

many subjects are reluctant to commit themselves to these longer durations, especially initially. This reluctance is more pronounced in subjects whom we classify as selfish, according to an independent measure of subjects' social types. Subjects classified as conditional cooperators, on the other hand, are more likely to self-select into commitment early on and throughout the experiment. Endogenous commitment results in significantly improved cooperation rates—in one of our conditions, the rate of cooperation is very close to 100 percent. These high cooperation rates are only partly explained by the differential sorting of selfish and conditionally cooperative types. Therefore, our results identify voluntary long-term commitment as a mechanism for facilitating cooperation, both by its differential appeal to types that differ in cooperativeness and through a more direct influence on behavior.

There are situations outside the laboratory in which interacting parties endogenously select the length of their relationship. For example, contractual relationships, as in employment, are usually for endogenously-specified periods of time. Similarly, professional partnerships and marriages generally involve lengthy or permanent voluntary commitment to one another by interacting parties. Of course, in the above situations, parties often have some information about each other prior to committing to long-term relationships, and they also often have the possibility of exiting the relationship, though perhaps at significant costs. However, instead of mimicking such concrete "real-world" situations in the laboratory, our design deliberately abstracts from these features to study the effect of long-term commitment, in isolation. That is, we used the PD as the stage game, a game that embodies a social dilemma in a simple, two-person, binary decision situation, and excluded any other confounding factors or complications such as endogenous partner choice, information about past actions, or the ability to exit relations prematurely. By stripping away everything else, we tried to create the cleanest and simplest test of the pure commitment effect on cooperation.

2. Theoretical Analysis

Our main interest lies in (1) investigating whether subjects opt for long-term commitment when it is available, and whether such commitment preferences differ by social type, and (2) comparing cooperation and earnings under long-term commitment with those under one-shot pairings and with those under exogenously imposed long-term commitment.

We consider an environment comprising both selfish players and other players who prefer to reciprocate cooperation. The basic intuition of the formal analysis below is that (1) everyone should opt for long-term commitment and (2) the set of equilibria is unchanged by the introduction of voluntary long-term interaction, relative to that under similar exogenously-imposed long-term commitment.[3] By choosing long-term interaction, conditional cooperators can protect themselves against continued exploitation by selfish players. That is, after being exploited once, the conditional cooperator learns the type of her opponent and has the possibility to avoid future losses by switching to defection, too. This is not possible in the one-shot option because there, the conditional cooperator faces a string of unknown opponents. Since conditional cooperators prefer long-term contracting, choosing the one-shot environment guarantees an encounter with a selfish player, meaning that selfish players cannot gain by attempting to exploit the one-shot interactions by defecting on conditional cooperators. Consequently, selfish players also universally select the long-term option. Such pooling leaves the set of equilibria in the game with voluntarily chosen long-term commitment identical to that when similarly long commitment is exogenously imposed. In particular, there exist behaviorally similar equilibria in which even selfish types cooperate "rationally" until the last periods of the repeated interaction. In addition, there also always exist equilibria with universal defection, in which case the choice of interaction duration is irrelevant.

To investigate which (pure-strategy) Perfect Bayesian Nash Equilibria exist when players can choose the duration of commitment to a randomly selected opponent, we consider a very simplified, "gang of four"-style model with heterogeneous players (Kreps

3. See the appendix for a more detailed exposition of the theoretical analysis and proofs that form the basis for this section.

et al. 1982). An interaction involves two players, drawn at random from the population, playing a generic PD. Table 1 shows players' payoffs. Payoffs for mutual cooperation and mutual defection are normalized to 1 and 0, respectively, and $r > 1 > 0 > q$ and $0 < r + q < 2$.[4]

We assume two types of players: selfish players and conditional cooperators. Both types maximize their undiscounted von Neumann-Morgenstern utilities, as given in Table 1. For the selfish type, these utilities coincide with the monetary payoffs of the game (1, 0, r, and q). Conditional cooperators (CC's), however, prefer to cooperate if the partner cooperates and prefer to defect if the partner defects. We model such preferences through a psychic cost that makes their temptation payoff, r', worse than the cooperative payoff, 1, so that $0 < r' < 1 < r$. Types are private information, but all players hold rational and commonly known beliefs about the distribution of types in the population.

Table 1: Payoffs in a generic PD. For conditional cooperators, the temptation payoff r is reduced to $r' < 1$.

		Player 2	
		C	D
Player 1	C	$1, 1$	$q, r^{(\prime)}$
	D	$r^{(\prime)}, q$	$0, 0$

2.1. Exogenous Interaction Horizons

We first consider the case in which the number of periods with a fixed opponent is exogenously imposed. Specifically, assume that there are T total periods of interactions, or repetitions of the game, but that each random pairing of players lasts for some fixed number of periods, $S \in \{1, 2, \ldots, T\}$. For simplicity, assume that $T = 2$, meaning that either $S = 1$ (i.e., players play in two one-shot interactions) or $S = 2$ (players are paired with the same counterpart in both periods of the game).

4. Colombo and Merzoni (2006, 2008) consider a bilateral relationship with endogenous length. However, their model uses a trust game, and they do not study the signaling and sorting properties of the equilibria.

Suppose first that $S = 1$. Then, we have a string of two one-shot PD interactions. In a one-shot game, a selfish player will always defect, regardless of her belief about the other player, because defection is the dominant strategy. The conditional cooperator will cooperate only if her belief, $0 \leq b \leq 1$, that the other player is also a conditional cooperator is sufficiently high, namely $b \geq {}^{-q}/_{1-r'-q}$. Since our focus is on cases in which the cooperative equilibria can arise, we assume that the share of conditional cooperators in the population is sufficiently high to allow this. Note, however, that defection is also an equilibrium strategy because even conditional cooperators defect if they believe other conditional cooperators will do so.[5] Thus, cooperation is possible whenever the frequency of conditional cooperators is sufficiently high, but universal defection always remains an equilibrium.

Next, suppose that $S = 2$, meaning that players are paired at random in the first period but then play twice with the same opponent. Because the second period is the last period in the repeated game, we have the same two equilibria as in the one-shot case, where cooperation can be supported only if $b \geq {}^{-q}/_{1-q-r'}$. However, in the first period, cooperation can now be supported as an equilibrium strategy for both types. Specifically, when $b \geq 1 - {}^{1}/_{r}$, an equilibrium exists in which both types of players cooperate initially, selfish players always defect in the second period ("rational cooperation"), and conditional cooperators use a trigger strategy, that is, they cooperate in the second period only if their opponent cooperated in the first period.

Finally, consider two settings with the same number of total periods, e.g., $T = 4$, and with one involving longer interactions ($S = 4$) than another ($S = 2$). Under the above cooperative equilibrium, the setting with $S = 4$ will yield higher overall cooperation rates and aggregate payoffs than the setting with $S = 2$, because the former contains just one end-game period, in which cooperation breaks down because the selfish types defect with certainty. This intuition highlights the first prediction for our experimental results.

5. If two conditional cooperators face each other, and their type is common knowledge, they are effectively playing a stag hunt coordination game.

Prediction 1 (Cooperation in the Repeated Game with Exogenous Horizons) *Holding the total number of periods constant, cooperation will be more frequent, and thus average payoffs will be higher, when players are exogenously matched for one longer interaction horizon rather than for multiple shorter ones.*

2.2. Endogenous Interaction Horizons

What happens if players can choose, *ex ante*, between a long-term commitment option, in which they are paired with the same opponent for multiple periods (e.g., a "commit" option, or $S = 2$) or a series of one-shot PD's with different opponents ($S = 1$)? Both the selfish players' behavior in the PD and players' commitment choices hinge on the cooperativeness of the CC's in the two commitment options. Recall that selfish players will always defect in the one-shot option, but in the "commit" option, they can be induced to cooperate rationally if the CC's cooperate conditionally. Thus, if CC's cooperate in neither of the two options, then the selfish players will also always defect; the commitment choice is then irrelevant as both options yield a certain payoff of zero.

However, equilibria that involve cooperation by CC's imply pooling. To see why, consider first that if CC's are cooperative in just one option, all players will choose this option. For instance, if CC's cooperate in the "no commit" ($S = 1$) option but defect in the "commit" ($S = 2$) option, the selfish players will still defect in both options. Moreover, both types will pool on "no commit" because, for the CC's, the prospect of encountering other cooperating CC's outweighs the loss from meeting defecting selfish players, and selfish players follow the cooperating CC's. On the other hand, if CC's cooperate conditionally in the "commit" option but defect in the one-shot option, this will induce the selfish players to cooperate rationally in the "commit" option. Both types will prefer this to mutual defection in the "no commit" option, resulting in pooling on "commit." Finally, if CC's cooperate (conditionally) in both options, the selfish players will cooperate rationally in the "commit" option. Since repeated interaction is more attractive for CC's in this case, they will commit, and the selfish players will follow. In general, CCs' payoffs are not only higher in expectation in the "commit" option, the variance of their

payoff is also lower, and these differences become more pronounced as the interaction horizon increases. We therefore expect that the equilibrium with pooling on the "commit" option, conditional cooperation by CC's, and rational cooperation by selfish players will be preferred.[6]

To summarize, if the equilibrium strategy of the CC's contains cooperation, then both types will pool in their choice of interaction horizon. Because beliefs over types remain unchanged if players pool on the same commitment option, this also means that the set of equilibria available in the repeated game with an exogenously imposed interaction horizon, S, is identical to when players endogenously choose this time horizon.

Prediction 2 (Pooling on Endogenous Interaction Horizons)
With endogenous choice of interaction horizons, there is no equilibrium where each type chooses a different commitment option, except under universal defection. Equilibrium cooperation when interaction horizon S is endogenously selected will coincide with that when the same horizon is imposed exogenously.

This prediction essentially states that the availability of endogenously selected interaction horizons will have no impact beyond simply making available the equilibria possible under that option, when exogenously imposed. This is because all equilibria—except ones with universal defection—involve pooling between the two types. Thus, beliefs about the composition of different types under a particular interaction horizon coincide with those when a particular horizon is exogenously imposed on the entire population.

2.3. Behavioral Prediction

The analysis above suggests that allowing endogenously chosen interaction durations should have little effect beyond when similar interaction durations are imposed exogenously. This is driven by pooling of conditional cooperators and selfish types in their commitment choices. However, one of our motivating intuitions is that long-term interaction may appeal differentially to different types. Therefore, we next attempt to provide

6. This is also in line with the experimental evidence on exogenous durations of the finitely-repeated PD, cited in the introduction, that repeated games yield higher payoffs than one-shot games.

a simple example, using a model of bounded rationality, of how such differential sorting may occur.

Specifically, above we assume that players possess rational beliefs about the proportions of the two types of players, and that their equilibrium beliefs about the proportions of players they encounter by selecting a particular interaction horizon are also accurate. Suppose, however, that players do not hold such rational beliefs, but instead possess heterogeneous beliefs about the rationality of their opponents, as in models of "level-k reasoning" or "cognitive hierarchies" (e.g., Nagel 1995; Costa-Gomes, Crawford, and Broseta 2001; Camerer, Ho, and Chong 2004; Ellingsen and Östling 2010). In this section, we show how a simple model of this variety can produce the prediction that different types select different contract durations. While while this analysis is simple and *ad hoc*, it captures the intuition motivating our central behavioral hypothesis.

Assume, as in our earlier discussion, that $T = 2$ and that there are two possible interaction horizons, $S = 1$ and $S = 2$, among which subjects choose. Our analysis above shows that, in any equilibrium in which conditional cooperators act cooperatively after selecting a particular interaction horizon, selfish types will follow the conditional cooperators, resulting in pooling. That is, if conditional cooperators select a long interaction horizon and subsequently cooperate, selfish players will also select the long interaction horizon.

Now assume instead, however, that players believe their opponents choose contracts at random (as in Level-0 behavior), and that the cooperative equilibrium obtains under either interaction horizon.[7] Because Level-0 players choose their commitment non-strategically, Level-1 players (who believe they face Level-0 opponents) will think that CC's are equally prevalent in both commitment options. It is straightforward to show that, in the above scenario, for players with Level-1 beliefs, the one-shot interaction will be more attractive for selfish types if and only if there are enough Level-0 conditional cooperators on whom they can defect. More precisely, if $b > 1/r$, then the expected share of conditional

7. That is, suppose that, among those who select $S = 1$, conditional cooperators cooperate and selfish types defect; among those who choose $S = 2$, both types cooperate in the first period and, in the second period selfish types defect while conditional cooperators reciprocate prior opponent behavior.

cooperators is high enough for the selfish types to choose the one-shot interaction. Level-1 conditional cooperators (who believe that the other players randomize) will always prefer $S = 2$ because, in the first period, they are guaranteed a payoff of 1. Thus, at level 1 behavior, only selfish types will choose the one-shot interaction.

Level-2 types, on the other hand, will recognize that there are more selfish types in the one-shot environment, and conversely expect higher proportions of conditional cooperators in the long horizon environment. They will therefore tend to choose the longer interaction horizon. Thus, these non-equilibrium beliefs are more likely to support cooperative equilibria in the endogenously chosen long interaction horizon, than when such a horizon is exogenously imposed on everyone. We therefore expect the prevalence of cooperative behavior to be higher in the longer time horizon than when a similar time horizon is exogenously imposed, due to the differential sorting of selfish and conditionally cooperative Level-1 types.

Prediction 3 (Limited Strategic Reasoning)
Due to limited strategic reasoning, we expect pooling to be imperfect initially. In particular, selfish types are likely to favor the one-shot option, relative to conditional cooperators. Cooperation will initially be higher in the endogenously chosen long interaction horizon.

3. Experiment Design

Table 2 shows the stage game Prisoner's Dilemma game (PD) studied in the experiment. The same game was used in two stages: a first stage in which we elicited behavioral (social) types and a second stage, constituting the main part of the experiment, in which the game was repeated 150 times with exogenously and endogenously varying interaction horizons.

3.1. Stage 1: Eliciting Behavioral Types

At the beginning of the experiment, after initial instructions, participants played a sequential PD one time. In this first stage, we used the strategy method to elicit a subject's

behavior both as a first mover and as a second mover. In section 4, we discuss how we construct social types based on the two choices made in the one-shot game.[8]

Subjects were randomly and anonymously paired. They then indicated a choice, A (= cooperation) or B (= defection), in case they would be selected as the first mover. Then, they indicated their choice as a second mover, conditional both on whether the first mover cooperated or defected.

Participants did not receive feedback from this stage until after the main part of the experiment. Specifically, at the end of the experimental session, the computer randomly chose one subject in each pair to be the first mover and the other to be the second mover, and implemented the corresponding choices. We kept this first PD identical across treatments and instructions for the main part of the experiment (stage 2) were distributed only after the sequential PD was over. We also made clear that subjects' choices, matching, and payoff in the first stage were inconsequential for and unrelated to the second stage.

Table 2: Payoffs in the experimental Prisoner's Dilemma (in ECUs)

		Player 2: A	Player 2: B
Player 1	A	40 / 40	65 / 5
	B	5 / 65	20 / 20

3.2. Stage 2: Repeated Prisoner's Dilemma Game

After the sequential PD, subjects played a total of 150 periods of a *simultaneous* PD with the same payoff matrix as in the first stage (see Table 2). In each session, subjects were grouped into matching groups of 14 or 16 subjects. Subjects only encountered other subjects from their own matching group.

8. In an otherwise unrelated experiment, Burks, Carpenter, and Goette (2009) also conduct a sequential PD to categorize social types in their but they only use second mover behavior for their classification. In contrast, we use both first- and second-mover behavior to construct our types.

The 150 periods were divided into 15 sets of 10 periods each. The meaning and relevance of these sets varied by treatment condition. Our two treatment dimensions are (1) how frequently subjects were re-matched and (2) whether the re-matching frequency was exogenous or whether subjects chose how frequently they would be rematched. Concerning the frequency of re-matching, subjects could be committed to the same opponent for only one period (denoted as "N" for "**N**o commitment"); for one 10-period set (denoted as "I" for "**I**ntermediate commitment"); or for all remaining periods of the experiment (denoted as "P" for "**P**ermanent commitment").

3.3. Conditions with Exogenous Interaction Horizons

We implemented three different conditions with exogenously determined interaction horizons in player matches. In these conditions, re-matching occurred at pre-specified and publicly known time intervals. In each of these *exogenous* (EXO) conditions, subjects received instructions that specified how regularly re-matching would occur, and stating clearly that re-matching would be random.

In the Exogenous, No Commitment (EXO-N) condition, subjects were randomly re-matched within their matching group after every period. Thus, subjects played in 150 1-period matches.

In the Exogenous, Intermediate Commitment (EXO-I) condition, subjects were randomly re-matched within their matching group after every 10-period set. Thus, subjects played 15 10-period matches.

Finally, in the Exogenous, Permanent Commitment (EXO-P) condition, subjects remained matched with the same partner for all 150 periods of stage 2. That is, they were matched with the same counterpart for the entire experiment. In this case, the pair was also the matching group.

3.4. Conditions with Endogenously Chosen Interaction Horizons

In addition, we conducted two conditions in which subjects could choose, *ex ante*, for how long they wanted to commit to play with the same partner.

In a CHOICE-NI session, subjects chose before each 10-period set between no commitment, i.e., random re-matching after every period of the set, or intermediate commitment, i.e., a fixed pairing for the duration of the 10-period set.

In a CHOICE-NIP session, subjects could additionally choose permanent commitment, which would match them with a randomly selected counterpart for the remainder of the session. Thus, for example, if selected at the beginning of the first set (period 1), this matching option would pair a subject with the same partner for all 150 periods of the stage. If selected at the beginning of the second set (in period 11), this option would match a subject with the same partner for all remaining 140 periods. Once a subject entered a permanent match, the pairing was irreversible. In this condition, subjects could also choose the intermediate (10-period) or no commitment (1-period) options, as in the CHOICE-NI condition.

In these two CHOICE conditions, after subjects expressed a preference for a particular interaction horizon, the computer formed pairs of subjects who had chosen the same commitment option (no commitment, intermediate commitment, or, in the CHOICE-NIP, permanent commitment). Because perfectly implementing subjects' commitment choices requires that even numbers of subjects choose each option, sometimes a subject's choice could not be implemented. To resolve such cases, we started with the longest commitment option available in each condition (intermediate commitment in CHOICE-NI and permanent commitment in CHOICE-NIP). The computer verified whether there were an even or odd number of subjects selecting that option. If the number was even, then the choices were implemented by randomly matching these subjects. If the number was odd, one subject making this selection was selected at random to be matched according to the next longest interaction horizon (no commitment in CHOICE-NI and intermediate commitment in CHOICE-NIP). In CHOICE-NI, this procedure guaranteed an even number of participants in each interaction horizon. In CHOICE-NIP, it was still possible that an odd number remained then in the intermediate commitment option, in which case the computer would randomly select a subject who had chosen intermediate commitment

to be matched according to the no commitment option.[9] This procedure was explained, clearly, in the instructions.

Importantly, in both CHOICE conditions, subjects chose the number of periods to be paired with a *randomly* assigned partner. That is, they did not choose the specific partner, nor did they know anything about the specific partner, other than that the person had (most likely) selected the same interaction horizon. This distinguishes our research from related research on endogenous partner choice in which players make matching decisions with information about prospective partners (see, for example, Hauk and Nagel 2001; Fujiwara-Greve and Okuno-Fujiwara 2009).

Table 3 summarizes our five treatment conditions. The check marks indicate the commitment options available in each condition, at the beginning of each ten-period set. The table also shows the number of subjects in each condition, and the number of independent matching groups.

Table 3: Summary of treatment conditions: interaction horizons available in each condition; number of subjects (N) and matching groups (k)

	Duration of match			N	k
	1 Period	10 Periods	All remaining		
EXO-N	✓	—	—	32	2
EXO-I	—	✓	—	58	4
EXO-P	—	—	✓	30	15
CHOICE-NI	✓	✓	—	71	5
CHOICE-NIP	✓	✓	✓	76	5
Total				267	31

9. We designed this mechanism to minimize instances where a decision could not be implemented. For subjects who chose the "no commitment" option, implementation probability was 100%, by default. For subjects who chose "intermediate commitment", the realized probability of this choice being implemented was 93% in CHOICE-NI and 94% in CHOICE-NIP. For subjects who chose "permanent commitment" in CHOICE-NIP, the probability of this choice being implemented was 93%.

3.5. Experimental Procedures

This experiment was conducted in English at the Department of Economics at the University of Zurich. The experiment was computerized, and participants made their decisions privately and anonymously.[10] 268 students from the University of Zurich and the Swiss Federal Institute of Technology in Zurich participated in the experiment.[11] In each session, only one of the treatment conditions was conducted, and each participant took part in only one treatment (between-subjects design). None of the participants were students in economics or psychology. A session lasted about 1.5 hours and the participants earned about 45 CHF on average. Instructions are reproduced in the appendix.

4. Results

We first present the results of the type elicitation, using the one-shot sequential PD. We then proceed to analyze cooperation in the second stage of the experiment, which is the main focus of our paper.

4.1. Types

We used the sequential PD in stage 1 of the experiment to identify social types and beliefs about others' types. According to their second-mover behavior, subjects were categorized as "selfish" (defect regardless of what the first mover chooses) or "conditional cooperators" (CC, i.e., reciprocate cooperation with cooperation and defection with defection). According to their first-mover behavior, both the selfish subjects and the conditional cooperators were classified as "optimists" if they cooperated as a first mover and "pessimists" if they defected. We choose this terminology because both selfish subjects and conditional cooperators only cooperate if they believe that there is a high probability

10. Recruitment was conducted using ORSEE (Greiner 2004). The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

11. One subject in the CHOICE-NI condition got sick during the experiment and had to leave her session before the end of the experiment. A research assistant who was unaware of the purpose of the study replaced this subject for periods 41 to 150. We exclude the data for this subject from the sample.

that the second mover is a conditional cooperator. So, in terms of the model we presented earlier, first-mover behavior is a binary indicator of a subject's belief, b. Finally, subjects who always cooperated—both as first and as second mover—are classified as "altruists", and the remaining subjects as "inconsistent".

Table 4 summarizes the classification. For example, the first row corresponds to subjects who always defected, and are therefore classified as "pessimistic and selfish" (pSE). This type comprises roughly half of our sample. Overall, selfish subjects (those who always defect as second movers) make up about 60 percent of the sample, conditional cooperators slightly more than one third.[12] A negligible proportion is either altruistic or inconsistent. Of the selfish subjects and the conditional cooperators, roughly two thirds are pessimistic (defect as first movers), and one third is optimistic. Interestingly, beliefs and social types are correlated: conditional cooperators are almost four times as likely to be optimistic as selfish subjects ($p < 0.01$, Chi-squared, $N = 257$).

Table 4: Classification of Social Types

1st-mover choice	2nd-mover response		Type	Frequency
	after cooperate	after defect		
defect	defect	defect	Pess. Selfish (pSE)	51.7 %
cooperate	defect	defect	Opt. Selfish (oSE)	9.7 %
defect	cooperate	defect	Pess. CC (pCC)	13.1 %
cooperate	cooperate	defect	Opt. CC (oCC)	21.7 %
cooperate	cooperate	cooperate	Altruist (Alt)	1.9 %
— all other combinations —			Inconsistent (Inc)	1.9 %

4.2. Exogenous Conditions

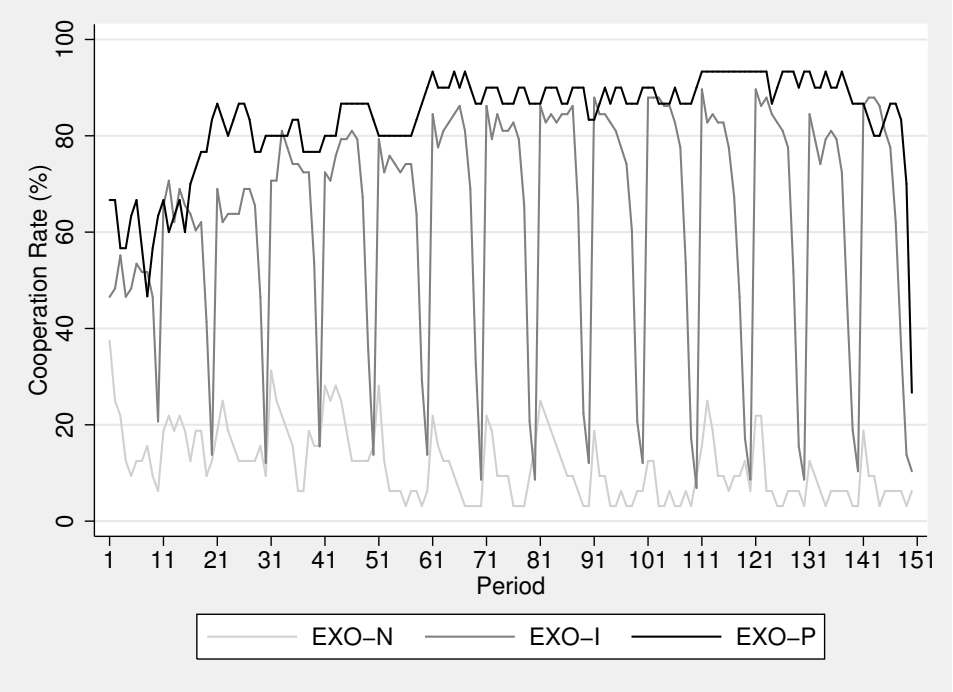
In Stage 2 of the experiment, subjects played 150 repetitions of the PD, in varying conditions. In our analysis, we put an emphasis on the first period, before subjects have received feedback for the first time. In this period, we can still treat each individual action

12. Comparing the type distribution across conditions, selfish and conditionally cooperative subjects make up the majority of subject types in all conditions. We fail to reject equality of proportions of types across conditions, indicating that randomization of types was successful ($p = 0.29$, Chi-squared, $N = 267$)

as an independent observation, which greatly increases statistical power. We begin by analyzing the three exogenous conditions: EXO-N, EXO-I, and EXO-P. Figure 1 shows the development of the average cooperation rate over time.

As expected, the average cooperation rate increases with the duration of the commitment. Cooperation and earnings are very low in the EXO-N condition, where subjects are rematched in every period. Despite slight increases at the beginning of every 10-period set, cooperation rates are rarely sustained above 10 percent, after the first few sets.

Figure 1: Cooperation Rate over Time in Exogenous Conditions



Cooperation rates in the two conditions with longer exogenously-imposed interaction horizons, EXO-I and EXO-P, are considerably higher. After the initial sets, cooperation rates are regularly above 80 percent. Cooperation is more frequent under permanent commitment (EXO-P), mainly due to the removal of the intermediate end games. That is, a large part of the difference between EXO-P and EXO-I is due to the decline in cooperation near the end of every 10-period match in the latter (a result similar to that obtained by Andreoni and Miller 1993).

Overall cooperation rates in the three conditions are: 11.5 percent (EXO-N), 63.5 percent (EXO-I), and 83.5 percent (EXO-P), see also Table 6. Using the matching group as the unit of observation, the distributions are significantly different ($p = 0.03$, Kruskal-Wallis Test, $N = 21$). Pairwise comparisons of the differences between conditions also generally reveal significant differences (EXO-N vs. EXO-I: $p = 0.06$, Mann-Whitney-U, $N = 6$; EXO-I vs. EXO-P: $p = 0.05$, Mann-Whitney-U, $N = 19$; EXO-N vs. EXO-P: $p = 0.05$, Mann-Whitney-U, $N = 17$).

As a direct consequence, earnings are also generally higher with longer exogenously imposed commitment. Specifically, the average profit per period in the EXO-N condition is 27.2 ECU versus 34.8 ECU in the EXO-I condition, and 37.7 ECUs in the EXO-P condition. Profit distributions are significantly different across these three conditions ($p = 0.03$, Kruskal-Wallis Test, $N = 21$). Differences between all pairs of conditions reflect similar significance levels to those for cooperation rates (EXO-N vs. EXO-I: $p = 0.06$, Mann-Whitney-U, $N = 6$, EXO-I vs. EXO-P: $p = 0.05$, Mann-Whitney-U, $N = 19$, EXO-N vs. EXO-P: $p = 0.05$, Mann-Whitney-U, $N = 17$).[13]

Result 1 (Cooperation and profits with exogenous long-term commitment) *On average, longer exogenously-imposed commitment increases cooperation and earnings. Earnings and cooperation are highest under permanent commitment and lowest under no commitment.*

Our classification of social types is meaningful in terms of predicting first-period behavior in EXO-N. While pessimistic selfish subjects overwhelmingly defect (15.8 percent cooperation), optimistic CC's predominantly cooperate (83.3 percent cooperation). The difference in cooperation rates is significant ($p < 0.01$, Chi-square, $N = 25$).[14] The classification is also predictive of first-period behavior in the two conditions with repeated-game incentives, EXO-I and EXO-P. In particular, optimistic CC's cooperate at a much higher rate in period 1 (90.9 percent) than the other three types—pessimistic (38.6) and

13. Again, in all of these comparisons, the unit of observation is the matching group.

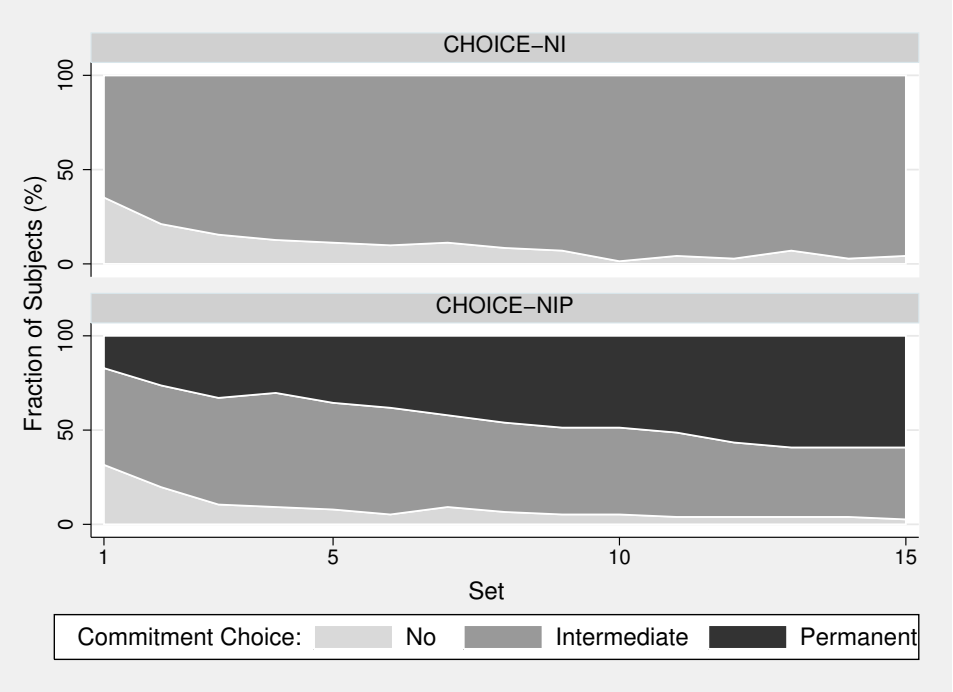
14. In EXO-N, there are only three subjects who are optimistic egoists and two who are pessimistic CC's, making statistical inference for these types infeasible.

optimistic (44.4) egoists, and pessimistic CC's (30.0). The latter three frequencies are statistically identical ($p = 80.4$, Chi-squared Test, $N = 63$), while the higher cooperation for optimistic CC's is statistically significant against all other three types ($p < 0.01$ for all three comparisons, Chi-squared Test, $N = 66$, $N = 31$, and $N = 32$, respectively).

4.3. Endogenous Conditions: Commitment Choice

In the endogenous conditions, subjects have to indicate how they want to be matched at the beginning of each 10-period set. In the CHOICE-NI condition, they can choose between being randomly rematched after every period (no commitment, as in EXO-N) or being matched with the same (randomly assigned) opponent for all 10 periods of the upcoming set (intermediate commitment, as in EXO-I).

Figure 2: Rate of commitment choices over Time



The first panel of Figure 2 shows the evolution of the matching choice distribution over time in the CHOICE-NI condition. At the beginning of the first set, i.e., in period 1, about 35 percent of subjects choose the no commitment option. The average rate of

subjects choosing this option drops permanently below 10 percent by set 8. Interestingly, however, a small number of subjects continue to opt for no commitment throughout the experiment.

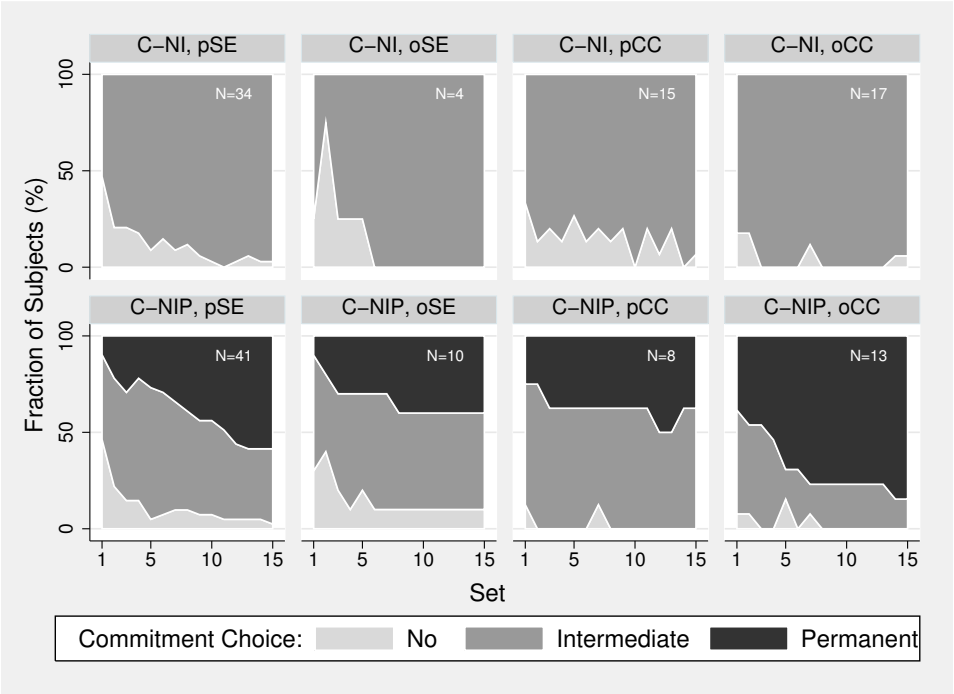
In CHOICE-NIP, subjects have the additional option to be *irreversibly* paired with a randomly assigned opponent ("permanent" commitment, as in EXO-P). The second panel of Figure 2 shows the distribution of commitment choices, over time, for this condition. We see that a very similar fraction of subjects initially chooses no commitment (32 percent), as in the CHOICE-NI condition. Similarly to CHOICE-NI, this fraction drops permanently below 10 percent by set 5 and below 5 percent by set 11. Again, at the end of the experiment, there remains a small residual proportion of subjects who persistently opt for no commitment.

Most interesting in this condition is how many subjects opt for permanent commitment. The fact that virtually everybody abandons no commitment in favor of some form of commitment suggests that subjects realize the advantage of committing. Also recall the significant benefits—discussed in our theoretical analysis and confirmed in the exogenous conditions—of permanent commitment, in particular the obvious lack of intermittent end-game defection. Nonetheless, subjects are very reluctant at first to choose this option. Only 17 percent of all subjects in CHOICE-NIP choose permanent commitment in the first set (one fourth of those who choose to commit). By set 5, this fraction increases to about 35 percent, and finally plateaus in the last three sets of the experiment at 59 percent. Of those who commit by the end of the experiment, 39 percent still do not commit permanently.

Result 2 (Fear of commitment)

About one-third of subjects opt for no commitment initially. Longer-term commitments increase in prevalence over time, driving out one-period matches. However, a large proportion, 41 *percent of subjects, persistently avoids permanent commitment.*

Figure 3: Commitment Rate by Condition and Type, over Sets



4.4. Which types opt for commitment?

We now consider whether different social types opt differentially for long-term commitment. Since there is a time trend in commitment choice, we first ask which type initially sorts into commitment, in period 1, and then consider who persistently avoids commitment throughout the experiment.

Figure 3 shows the proportion of subjects opting for commitment over time, presented separately for each subject type and condition. these graphs are analogous to those in Figure 2, but with the commitment choices of each type presented separately.

In the first set, commitment is highest for optimistic conditional cooperators (oCC) and lowest among pessimistic self-interested (pSE) types, and this is true for both conditions. Only half of the pessimistic selfish subjects opt for commitment in either condition. At the other extreme, 82 percent of optimistic CC's opt for commitment in the first period of the CHOICE-NI condition; in the CHOICE-NIP condition, the fraction is even higher, 92 percent. This is in line with our hypothesis that selfish subjects are more inclined

to initially opt for one-period matches, while conditional cooperators favor long-term commitment.

The possibility to commit irreversibly in the CHOICE-NIP condition has no effect on the initial commitment rate of selfish subjects, but makes CC's commit more often: pessimistic CC's are 20 percentage points more likely to choose some kind of commitment in CHOICE-NIP than in CHOICE-NI, and optimistic CC's are 10 percentage points more likely. When permanent commitment is available, over 90 percent of optimistic CC's (i. e., those who are most likely to cooperate) opt for either intermediate or permanent commitment from the start.

In Table 5, we test the relationship between types and commitment choices econometrically, by regressing the decision to opt for some kind of long-term commitment—intermediate or permanent—on binary variables indicating a subject's type, a binary variable for the CHOICE-NIP condition, interaction terms, and control variables. Looking at model 2, optimistic CC's commit significantly more frequently than pessimistic selfish subjects (the excluded category) in both the CHOICE-NI and the CHOICE-NIP conditions. Specifically, they are 29 percentage points more likely to commit in the CHOICE-NI condition ($p = 0.04$) and 37 percentage points more likely in CHOICE-NIP ($p = 0.01$ in a Wald test). As mentioned above, optimistic CC's are about 9 percentage points more likely to commit in CHOICE-NIP than in CHOICE-NI, but this is not statistically significant. However, while the CHOICE-NIP condition does not make any individual type significantly more likely to commit than in the CHOICE-NI condition (i.e., none of the CHOICE-NIP terms is statistically significant), the pessimistic CC's are now also more likely to commit than pessimistic selfish subjects—specifically, 33 percentage points more likely and this difference is statistically significant ($p = 0.06$, Wald test). As model 3 shows, these results are robust to adding demographic controls (gender, age and Swiss nationality), which are all insignificant, and controls that capture subjects' risk and trust attitudes.[15]

15. Of the eight trust questions, adopted from Glaeser et al. (2000), two are (marginally) significant ("Trustworthiness", $p = 0.05$ and "GSS Help", $p = 0.09$); of the seven risk-related questions, taken from Dohmen et al. (2011), three are significant: risk taking in sports and leisure (positively correlated, $p = 0.01$), risk taking in professional career (negatively correlated, $p < 0.01$), and the amount invested

Figure 3 also shows the development of commitment over time. At the end of the experiment, optimistic CC's are those who have chosen permanent commitment most often (84.6 percent) in CHOICE-NIP. Intriguingly, pessimistic selfish subjects are by the end more likely to have committed permanently (58.5) than pessimistic CC's (37.5) and optimistic selfish types (40).[16]

Result 3 (Optimistic conditional cooperators favor commitment)
Optimistic conditional cooperators initially choose commitment with the highest frequency in either condition with endogenous commitment choices. Toward the end of the experiment, optimistic conditional cooperators are the most likely to have opted for permanent commitment.

These differences in the propensity to commit affect the type composition of subjects interacting under the different commitment choices. That is, if one chooses commitment, one is likely to encounter different proportions of subject types than in the population as a whole. This is inconsistent with the pooling equilibria we identified earlier, and provides a foundation for why the level of beliefs, b, necessary to support cooperation may be more likely to be met under endogenously selected interaction horizons than when everyone is assigned to a particular interaction duration. Specifically, as we hypothesized, the proportion of cooperative subjects is likely to be higher under endogenously selected commitment, relative to exogenously imposed commitment and to no commitment. This makes cooperation more likely under voluntarily selected long-term commitment, relative to when similar commitment is exogenously imposed, as the necessary optimistic beliefs are more likely to be met.

Figure 4 shows the proportion of the population comprised by each type, in each commitment category in the two conditions with endogenous interaction horizons. The figure presents the proportions separately for the first 10-period set and in the final

in a hypothetical lottery (positively correlated, $p = 0.02$). We also include scores from both the Mach-IV machiavellism inventory and the Cognitive Reflection Test; both are insignificant.

16. Note however that, after the first period, subjects have interacted with each other, so that our unit of independent observation is now the matching group. Since we only have five matching groups of condition CHOICE-NIP, the number of independent observations too small for statistical comparisons.

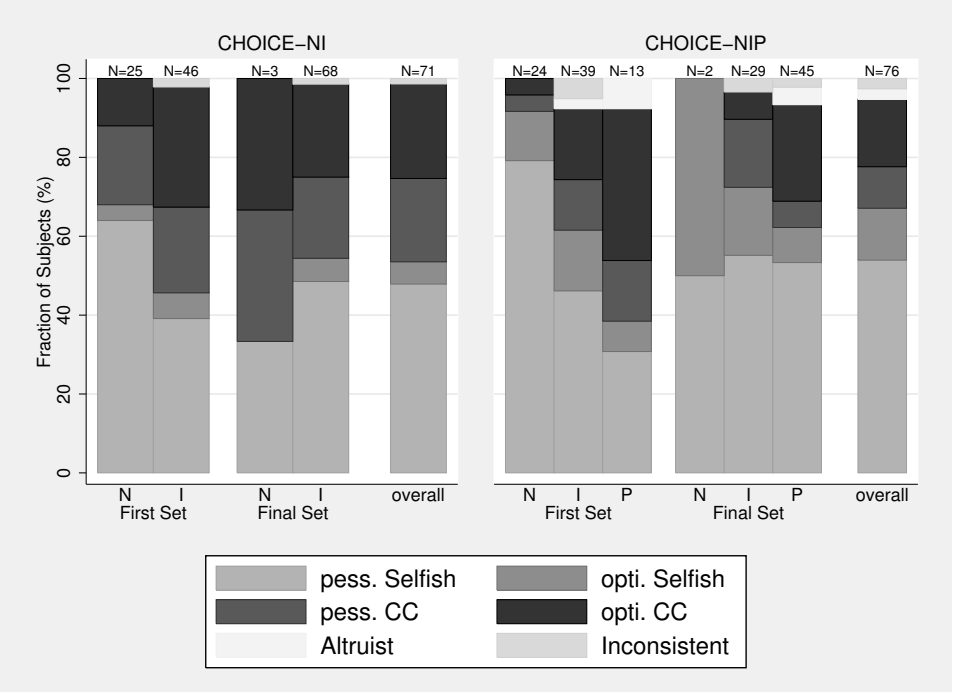
Table 5: OLS Regression of Commitment Choice in Period 1. Outcome is 0 if no commitment, 1 otherwise.

	M1	M2	M3
oSE	0.181	0.221	0.342
	(0.134)	(0.246)	(0.270)
pCC	0.206*	0.137	0.208
	(0.110)	(0.144)	(0.148)
oCC	0.333***	0.294**	0.361**
	(0.100)	(0.138)	(0.144)
CHOICE-NIP		0.007	0.015
		(0.108)	(0.111)
oSExCHOICE-NIP		-0.057	-0.277
		(0.296)	(0.321)
pCCxCHOICE-NIP		0.201	0.178
		(0.231)	(0.232)
oCCxCHOICE-NIP		0.092	-0.025
		(0.203)	(0.207)
Female			0.065
			(0.100)
Age >= median			0.080
			(0.090)
Swiss nationality			0.028
			(0.083)
Survey controls?	No	No	Yes
Constant	0.533***	0.529***	0.199
	(0.053)	(0.080)	(0.230)
adj. R-squared	0.063	0.045	0.108
N	142	142	142

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. In M3, the following controls from the exit questionnaire were added: attitudinal measures for trust (8 items from Glaeser et al. (2000)) and risk (7 items, from Dohmen et al. (2011)), a machiavellism score (from the Mach-IV inventory), and the score from the Cognitive Reflection Test (CRT).

one. As a reference, we show the overall distribution of types in the respective condition (labelled as "overall").

Figure 4: Type by Endogenously Chosen Interaction Horizon, First and Last Set



In the first set of both conditions, a subject is much more likely to encounter a selfish opponent when selecting the no commitment option than with longer commitment. For example, in the CHOICE-NI condition, selfish subjects comprise 68 percent of the population under no commitment, but only 46 percent under intermediate commitment ($p = 0.09$, Fisher's exact test, $N = 71$). A similar pattern obtains in the CHOICE-NIP condition. In the first set, a choice of no commitment virtually guarantees that one encounters a selfish counterpart (92 percent). With greater commitment, one encounters increasing proportions of conditional cooperators, and especially optimistic CC's, which make up a negligible proportion of the no commitment population, 18 percent of the population under intermediate commitment and 38 percent of the population under permanent commitment. Overall, one is much more likely to encounter a conditional cooperator under permanent commitment (54 percent) than under no commitment (8

percent, difference significant, $p = 0.01$, Fisher's exact Test, $N = 76$).

Finally, we consider commitment choices near the end of the experiment. As we saw above, by the end of the experiment, almost nobody remains in the no commitment option (3 out of 71 subjects in CHOICE-NI and 2 out of 76 subjects in CHOICE-NIP). However, in the CHOICE-NIP condition, almost 40 percent of subjects never opt for permanent commitment. Figure 3 shows that, by the final set, the two commitment choices, Intermediate and Permanent, are more similar in the overall distribution of types than in the first set. However, the fraction of pessimistic CC's is much larger in the Intermediate commitment option (17 percent) than in the Permanent option (7 percent); optimistic CC's, on the other hand, are much more prevalent in the Permanent option (24 percent versus 7 percent in Intermediate).

4.5. Endogenous Conditions: Cooperation Rate

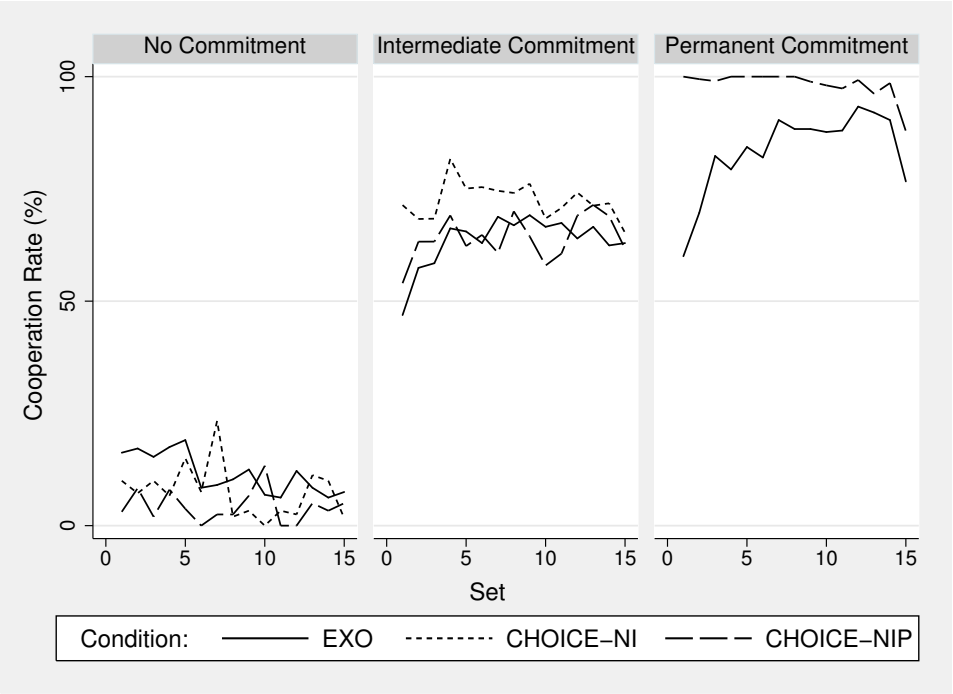
We have shown that, initially, subjects are more likely to encounter a conditional cooperator if they choose to commit, and in CHOICE-NI, even more if they choose to commit permanently. We also know that conditional cooperators, especially the optimistic ones, are more likely to cooperate initially. This means that, potentially, the longer-term commitment options facilitate cooperation when chosen endogenously, as they yield both a greater proportion of conditional cooperators and thus a greater proportion of counterparts likely to cooperate. That is, the endogenously chosen long-term commitment offers more than the repeated game incentive to cooperate. It also produces a type composition conducive to cooperation, due to a higher proportion of optimistic conditional cooperators. This is particularly true early in the experiment, when the differential selection of types into long-term contracts is most pronounced.[17]

17. Recall that our matching mechanism ensured that everybody in the highest commitment option ("intermediate" in CHOICE-NI and "permanent" in CHOICE-NIP) was matched with somebody who chose this option. Also, a subject whose choice was not implemented was guaranteed to be matched with somebody who chose the implemented option. Finally, in any given set, somebody who chose no commitment had a substantial chance of being matched with a participant who had been reallocated to no commitment from intermediate commitment by the computer mechanism (46% in CHOICE-NI and 24.6% in CHOICE-NIP); somebody who chose intermediate commitment in CHOICE-NIP had a 6.7% chance of meeting somebody who had chosen permanent commitment but was reallocated to

Table 6: Average Cooperation Rate (in percent) across commitment durations: exogenous vs. endogenous commitment

	1st Period of First Set				1st Period of Last Set				All Periods			
	N	I	P	aggr.	N	I	P	aggr.	N	I	P	aggr.
EXO	37.5	46.6	66.7	—	18.8	86.2	86.7	—	11.5	63.5	83.5	—
C-NI	14.3	76.7	—	52.1	16.7	89.2	—	83.1	9.0	72.5	—	63.5
C-NIP	7.7	37.5	100.0	35.5	25.0	83.3	100.0	89.5	4.5	63.9	97.8	70.7

Figure 5: Cooperation Rate over time across commitment durations: exogenous vs. endogenous



The left part of Table 6 presents the cooperation rates in period 1 in the different treatment conditions, divided by commitment choice for the endogenous conditions.[18] Compared to the cooperation rate in exogenously assigned one-shot interactions (38 percent), the rate is lower under no commitment in both the CHOICE-NI (14 percent) and the CHOICE-NIP (8 percent) conditions ($p = 0.02$, Fisher's exact test, $N = 86$). This means that if subjects have the chance to opt out of the one-shot interaction, this dramatically decreases the cooperation rate among those who nevertheless choose this no commitment option.

In contrast, if subjects have voluntarily chosen to commit irreversibly (in CHOICE-NIP), they *all* cooperate, while subjects who were exogenously assigned to this option (EXO-P) only cooperate at a rate of 67 percent ($p = 0.04$, Fisher's exact test, $N = 40$). Thus, putting the above two results together, we find strong support for the hypothesis that endogenously chosen long-term commitment yields higher cooperation that when similar commitment is exogenously imposed.

Interestingly, the fact that more cooperative types sort into permanent commitment only accounts for part of the above differences in cooperation rates: if we regress cooperation on treatment condition and commitment duration dummies, the difference between endogenous and exogenous commitment is unaffected by controls for social type (see Table 7).[19] We take this as an indication that voluntary permanent commitment affects behavior not only by yielding different type compositions, but also by affecting subjects' beliefs.[20]

Interestingly, there is no monotonic relationship for intermediate commitment. If sub-

---

intermediate commitment by the computer. Subjects whose commitment choice was not implemented differed from those whose choice was implemented in their cooperation rates in the PD (see appendix).

18. The cooperation rates of the three exogenous conditions—EXO-N, EXO-I, and EXO-P—are the same as in Period 1 of Figure 1, which shows cooperation rates across conditions and interaction horizons over time.

19. The null hypothesis whether the coefficients EXO-P and CNIP-P are identical can be rejected both without and with controls for social type and survey measures of risk, trust, and machiavellism ($p = 0.05$ in M1 and $p = 0.07$ in M2, Wald tests). Results are robust to including the CRT score. Also, the dependent variable is average cooperation within a set, but results are unchanged if we do not aggregate, or if we take only first periods of each set.

20. One possible interpretation is that endogenously chosen long-term interaction facilitates equilibrium selection on cooperative equilibria, in the manner of forward induction.

Table 7: OLS Regression of Cooperation in the exogenous and Choice-NIP treatment conditions, controlling for subjects' social type. Dependent variable: individual cooperation rate within a set.

	M1	M2
EXO-I	0.527***	0.517***
	(0.061)	(0.061)
EXO-P	0.715***	0.719***
	(0.081)	(0.081)
CNIP-N	-0.063*	-0.043
	(0.034)	(0.049)
CNIP-I	0.536***	0.547***
	(0.034)	(0.044)
CNIP-P	0.871***	0.850***
	(0.032)	(0.045)
oSE		0.059
		(0.041)
pCC		0.069*
		(0.038)
oCC		0.108***
		(0.033)
Survey controls?	No	Yes
Constant	0.107***	0.007
	(0.030)	(0.068)
adj. R-squared	0.546	0.567
N	2805	2805

Notes: * p<0.10, ** p<0.05, *** p<0.01. Standard errors account for clustering at the matching-group level. In M2, the survey controls are: trust and risk attitude, and machiavellism score.

jects are exogenously assigned to play the first set with the same opponent (EXO-I condition), 47 percent of them cooperate. When the 10-period commitment option is the only available commitment option (in CHOICE-NI), this rate jumps to 77 percent ($p < 0.01$, Fisher's exact test, $N = 101$). But when permanent commitment is also available (in CHOICE-NIP), cooperation under intermediate commitment actually decreases compared to EXO-I, to only 38 percent ($p < 0.01$, Fisher's exact test, $N = 83$). Thus, the cooperation-enhancing effect of commitment seems to be confined to cases in which players select the highest available commitment option.

Result 4 (Voluntary Long-Term Commitment Increases Cooperation) *Voluntarily choosing the longest-term commitment option available increases cooperation relative to when similar commitment is exogenously imposed. Conversely, voluntarily choosing no commitment decreases cooperation.*

5. Conclusion

In this paper, we add to the literature that investigates the influence of the shadow of the future on cooperation. In particular, we study how cooperation is influenced by endogenously chosen interaction durations, when individuals know nothing about their opponents beyond the kind of commitment choices that they made.

We find that, consistent with prior research, longer interaction horizons for an interaction facilitate cooperation. Moreover, the longer the interaction the greater the prevalence of cooperative behavior.

Importantly, we show that individuals realize these benefits of commitment and use the opportunity to lock into long-term interactions. By the end of the experiment, almost all participants choose some form of commitment.

However, we find at least two ways in which players' reluctance to commit limits some of the potential benefits produced by long-term interaction. First, we find that, initially, over a third of subjects opt for no commitment, meaning that they interact with a different counterpart every period, a situation very likely to yield mutual defection and low payoffs.

Second, even though the willingness to be bound in long-term interaction increases over time, we find that permanent commitment, which yields the highest cooperation rates, is avoided by a persistently large proportion of subjects.

We also find that long-term commitment appeals differentially to different types, in terms of social types that we identify separately from the main part of the experiment. In early periods, commitment is spearheaded mostly by optimistic conditional cooperators. In contrast to other studies where the influx of selfish subjects makes cooperation break down (e.g., Bohnet and Kübler 2005), we find that selfish subjects who choose long-term commitment "behave", a sign that subjects achieve coordination on the rational-cooperation equilibrium. So although conditional cooperators were in the minority in our sample, their presence and the possibility to choose interaction durations was sufficient to tip the scales in favor of commitment and cooperation.[21] Optimistic conditional cooperators are also most likely to commit permanently by the end of the experiment.

The persistent avoidance of commitment in our study is puzzling. As subjects realize the merits of intermediate commitment in fostering cooperation, they should prefer the permanent commitment *all the more*. However, while subjects move away from no commitment to some commitment, many are reluctant to give up the possibility to withdraw from commitment in the future, even if this comes both at a loss in expected payoff and increased payoff variance. Such "fear of commitment" has also been studied in psychology (Serling and Betz 1990), although there it describes how people adhere to action plans, rather than relationships. Bowlby (1969) studies a phenomenon called "attachment style", a classification of people's emotions and behavior with respect to the parent-child relation and romantic relationships. Numerous questionnaire studies have found that people are heterogeneous with respect to their attitude to attachment ranging from "securely attached" to "avoidant" (for an overview, see Cassidy and Shaver (2008)). Whether the fear of commitment observed in our experiment is related to such personality

21. Prior research has shown that whether one or the other type dominates aggregate outcomes is dictated by the relevant institutions and norms in a particular setting (Roth et al. 1991; Fehr and Falk 1999; Henrich et al. 2001; Camerer and Fehr 2006). Our results suggest that enforceable commitment to long-term interaction can be such an institution that helps the conditional cooperators shape aggregate outcomes.

traits remains to be investigated in future studies, as this could help identify its impact in other economic contexts, such as labor markets and organizations.

Finally, one of our most important findings is that self-selection into a commitment option has a significant effect on cooperation rates, especially initially. Those who actively sort into one-shot matching cooperate significantly less than those who are exogenously assigned into this option. Those who choose the longest possible commitment option cooperate significantly more. The permanent commitment option, when chosen endogenously, yields essentially *universal* cooperation. Importantly, sorting by type explains only part of this gap, which suggests that people may use permanent commitment as a device to coordinate efficiently on the cooperative equilibrium. Moreover, a comparison between our CHOICE-NI and CHOICE-NIP conditions shows that once intermediate commitment is not the longest possible commitment, it looses its value as a cooperation-fostering institution. The willingness to commit for the longest available horizon thus also serves as a way to screen the opponent for cooperativeness.

This goes beyond what prior research has shown. Vikander (2013) argued that in finitely-repeated social dilemmas, the best way to sustain cooperation is through sorting, whereby conditional cooperators manage to identify each other and to isolate themselves from opportunistic, selfish people. We show that, while sorting plays a role, endogenous commitment choice gives a large, lasting boost to cooperation even once almost everybody chooses it and the sorting argument is thus of minor importance.

To our knowledge, this paper is the first to isolate and investigate the cooperation-enhancing effect of voluntary commitment to long-term interaction. This sort of contractually enforceable commitment is a defining element of firms, organizations and social institutions. Because the self-selection effect creates higher levels of cooperation than even the exogenously induced long-term commitment, we think that voluntary commitment is a powerful advantage of organizations and can potentially explain their existence, in addition to more standard accounts like governance improvements. The large effect of irreversible commitment on the willingness to cooperate may be one reason why many cultures historically encourage long-term commitment between adults, in the form

of marriage (Phillips 1991). If marriage is a (quasi-) irreversible decision, it can serve both as a commitment device to cooperation and a signaling device to credibly convey trustworthiness (Matouschek and Rasul 2008, and references there).[22]

The endogenous choice of longer-term commitment may be one way to achieve "good governance" in groups with the aim to uphold cooperation (Dixit 2009). While irreversible commitment achieves very high cooperation rates, it is not taken up by everybody. Further work should investigate whether contract types that offer more flexibility are more attractive. Such a contract could be a commitment contract of indefinite duration that allows unilateral termination at a penalty. While this construct seems more realistic than the irreversible commitment in our experiment, it may still constitute a sufficient commitment to ensure a high level of cooperation.

References

Andreoni, James, and John H. Miller. 1993. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *The Economic Journal* 103 (418): 570–585.

Bohnet, Iris, and Dorothea Kübler. 2005. "Compensating the cooperators: is sorting in the prisoner's dilemma possible?" *Journal of Economic Behavior & Organization* 56 (1): 61–76.

Bowlby, John. 1969. *Attachment and Loss, Volume I: Attachment.* London: Hogarth Press.

Brown, Susan L., and Alan Booth. 1996. "Cohabitation versus Marriage: A Comparison of Relationship Quality." *Journal of Marriage and Family* 58 (3): 668–678.

22. Consistent with this view, married couples are significantly happier than cohabiting couples (Nock 1995; Brown and Booth 1996; Stack and Eshleman 1998).

Burks, Stephen, Jeffrey Carpenter, and Lorenz Goette. 2009. "Performance pay and worker cooperation: Evidence from an artefactual field experiment." *Journal of Economic Behavior & Organization* 70 (3): 458–469.

Camera, Gabriele, and Marco Casari. 2009. "Cooperation among Strangers under the Shadow of the Future." *The American Economic Review* 99 (3): 979–1005.

Camerer, Colin F., and Ernst Fehr. 2006. "When Does "Economic Man" Dominate Social Behavior?" *Science* 311 (5757): 47–52.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *The Quarterly Journal of Economics* 119 (3): 861–898.

Cassidy, Jude, and Phillip R Shaver. 2008. *Handbook of attachment: Theory, research, and clinical applications.* Guilford Press.

Colombo, Ferdinando, and Guido Merzoni. 2006. "In praise of rigidity: The bright side of long-term contracts in repeated trust games." *Journal of Economic Behavior & Organization* 59 (3): 349–373.

———. 2008. "For how long to tie your hands? Stable relationships in an unstable environment." *Journal of Economics* 95 (2): 93–120.

Costa-Gomes, Miguel, Vincent P. Crawford, and Bruno Broseta. 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica* 69 (5): 1193–1235.

Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich. 2008. "The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures." *The American Economic Review* 98 (4): 1443–1458.

Crawford, Vincent P., and Nagore Iriberri. 2007. "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental "Hide-and-Seek" Games." *The American Economic Review* 97 (5): 1731–1750.

Dal Bó, Pedro. 2005. "Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games." *The American Economic Review* 95 (5): 1591–1604.

Dixit, Avinash. 2009. "Governance Institutions and Economic Activity." *The American Economic Review* 99 (1): 3–24.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the European Economic Association* 9 (3): 522–550.

Ellingsen, Tore, and Robert Östling. 2010. "When Does Communication Improve Coordination?" *The American Economic Review* 100 (4): 1695–1724.

Fehr, Ernst, and Armin Falk. 1999. "Wage Rigidity in a Competitive Incomplete Contract Market." *Journal of Political Economy* 107 (1): 106–134.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10 (2): 171–178.

Friedman, James W. 1971. "A Non-cooperative Equilibrium for Supergames." *The Review of Economic Studies* 38 (1): 1–12.

Fujiwara-Greve, Takako, and Masahiro Okuno-Fujiwara. 2009. "Voluntarily Separable Repeated Prisoner's Dilemma." *Review of Economic Studies* 76 (3): 993–1021.

Gächter, Simon, and Armin Falk. 2002. "Reputation and Reciprocity: Consequences for the Labour Relation." *Scandinavian Journal of Economics* 104 (1): 1–26.

Ghosh, Parikshit, and Debraj Ray. 1996. "Cooperation in Community Interaction Without Information Flows." *The Review of Economic Studies* 63 (3): 491–519.

Glaeser, Edward L., David I. Laibson, José A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." *The Quarterly Journal of Economics* 115 (3): 811–846.

Greiner, Ben. 2004. "The Online Recruitment System ORSEE – A Guide for the Organization of Experimens in Economics." In *Forschung und wissenschaftliches Rechnen,* edited by Kurt Kremer and Volker Macho, 79–94. GWDG-Berichte 63. Göttingen: Gesellschaft für wissenschaftliche Datenverarbeitung.

Hauk, Esther, and Rosemarie Nagel. 2001. "Choice of Partners in Multiple Two-Person Prisoner's Dilemma Games." *Journal of Conflict Resolution* 45 (6): 770–793.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *The American Economic Review* 91 (2): 73–78.

Jackson, Matthew O., and Alison Watts. 2010. "Social games: Matching and the play of finitely repeated games." *Games and Economic Behavior* 70 (1): 170–191.

Kreps, David M, Paul Milgrom, John Roberts, and Robert Wilson. 1982. "Rational cooperation in the finitely repeated prisoners' dilemma." *Journal of Economic Theory* 27 (2): 245–252.

Matouschek, Niko, and Imran Rasul. 2008. "The Economics of the Marriage Contract: Theories and Evidence." *The Journal of Law and Economics* 51 (1): 59–110.

Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental Study." *The American Economic Review* 85 (5): 1313–1326.

Nock, Steven L. 1995. "A Comparison of Marriages and Cohabiting Relationships." *Journal of Family Issues* 16 (1): 53–76.

Page, Talbot, Louis Putterman, and Bulent Unel. 2005. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency." *The Economic Journal* 115 (506): 1032–1053.

Phillips, Roderick. 1991. *Untying the Knot: A Short History of Divorce.* Cambridge (UK): Cambridge University Press. ISBN: 9780521423700.

Rob, Rafael, and Huanxing Yang. 2010. "Long-term relationships as safeguards." *Economic Theory* 43 (2): 143–166.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *The American Economic Review* 81 (5): 1068–1095.

Serling, Deborah A., and Nancy E. Betz. 1990. "Development and Evaluation of a Measure of Fear of Commitment." *Journal of Counseling Psychology January 1990* 37 (1): 91–97.

Stack, Steven, and J. Ross Eshleman. 1998. "Marital Status and Happiness: A 17-Nation Study." *Journal of Marriage and Family* 60 (2): 527–536.

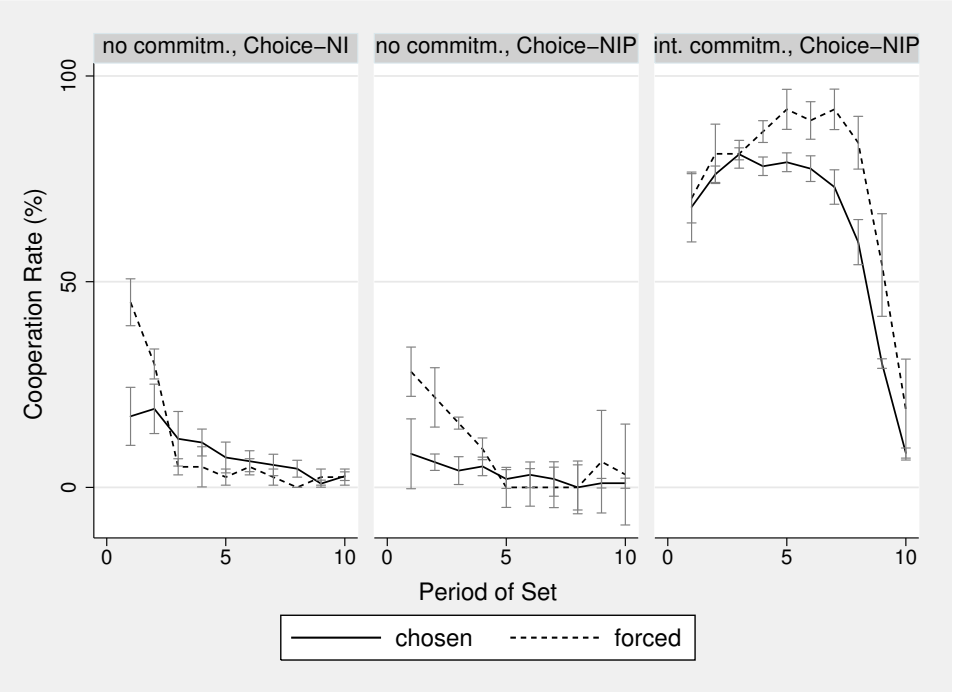
Vikander, Nick. 2013. "Sorting and sustaining cooperation." *Oxford Economic Papers* 65 (2): 548–566.

Appendices

A. Cooperation Rate when Choice was not implemented

The computer mechanism that implements subjects' commitment choices in the endogenous treatments may sometimes reassign a subject in case there is an uneven number of subjects in the chosen option. In Figure 6, we show the differential cooperation rates in the different commitment options of subjects who had chosen this option and those who wanted to be in the next higher option but were reallocated to this option by the computer. Those who were reallocated to "no commitment" display, initially, higher cooperation rates. After having experienced the low level of cooperation in this commitment option for a few periods, they reduce their cooperation to the low level of those who intended

Figure 6: Cooperation rate within a set in endogenous conditions. Solid lines show cooperation rate of subjects who chose the respective option; dashed lines show the commitment rate of subjects who were reallocated to the option by the implementation mechanism. Error bars indicate the standard error of the mean, with the matching group as the unit of observation.



not to be committed. In the intermediate commitment, subjects who had initially chosen "permanent" commitment sustain a higher cooperation rate over a longer period of time.

B. Theoretical Analysis

In this section, we provide details of the formal analysis summarized in Section 2. We study a signaling game consisting of the commitment choice ("Stage 1") and an entailing T-period Prisoner's Dilemma (PD), with or without commitment ("Stage 2"). We consider a model with two types of players, as described in Section 2: *selfish* players who maximize their monetary payoffs from the game and *conditional cooperators* who want to cooperate with those who also cooperate in return. A player's type is private information.

First, we provide a proof that, if the fraction of CC's is high enough, then any cooper-

ative equilibrium must entail pooling of types into commitment options. Then, we turn to a simpler version of this game to illustrate our motivating intuition in the form of a behavioral prediction.

To summarize the analysis, our first main result is that, under a reasonable behavioral assumption, all pure-strategy Perfect Bayesian equilibria of the supergame involving cooperative behavior—i.e., where commitment choices are not made irrelevant by universal defection—entail pooling by selfish and conditional cooperative types on the same commitment choice. Intuitively, it is impossible for conditional cooperators to select one type of interaction duration and cooperate, without the selfish types also choosing the same interaction duration.

However, our intuition motivating the experiment was based on the belief that conditional cooperators would be able to use the commitment choice to avoid selfish types, at least to some extent, and that the selfish types would have a preference for not committing, so they could exploit cooperation by others more easily. Therefore, we also show that it is possible, using a simple model of bounded rationality—the "level-k" approach of limited strategic reasoning—to construct a behavioral prediction in which (Level-1) conditional cooperators opt for longer commitment while (Level-1) selfish players do not. This provides one possible theoretical account for the behavior we observe in the experiment.

B.1. No Stable Separation of Types

In this section, we consider a finitely-repeated PD with two types of players, selfish and conditionally cooperative. Let T be the finite, commonly known number of periods in the game. Let $a_t \in \{D, C\}$ be an action in period t of the PD, where D stands for "defect" and C stands for "cooperate". The period payoffs are given in Table 1. Payoffs for mutual cooperation and mutual defection are normalized to 1 and 0, respectively, and $r > 1 > 0 > q$ and $0 < r + q < 2$.

We assume two types of players: selfish players and conditional cooperators. All players maximize their undiscounted expected von Neumann-Morgenstern utilities, as given in

Table 1. Selfish players realize a temptation payoff of $r > 1$. Conditional cooperators (CC's) prefer to cooperate if the partner cooperates and prefer to defect if the partner defects. We model such preferences through a psychic cost that makes the CCs' temptation payoff, r', worse than the cooperative payoff, 1, so that $0 < r' < 1 < r$.

The population of players is a continuum of non-atomistic players on the interval $[0, 1]$, where a fraction b is of the conditionally cooperative type and the complement, $1 - b$ is selfish. Types are private information, but all players hold the rational and commonly known belief b about the distribution of types in the population. In particular, this means that if a random opponent is drawn from this population, the probability of this opponent being a CC is b, and this fact is commonly known by all players.

For simplicity, we restrict our attention to pure strategies.[23] Players also recall their personal history, h_t^i, that is own play (a_τ^i) and opponent's play (a_τ^{opp}) for each of the previous periods, $\tau \in 1, 2, \ldots t-1$. In each period, player i can condition his strategy on the belief b and his personal history up to this period. Player i's strategy for the T-period game is then the vector $s^i = \{a_1^i(b), a_2^i(b, h_2^i), \ldots, a_T^i(b, h_T^i)\} \in S$.

In our model, there are two different matching protocols, no commitment (nc) and commitment (c). If a player is in the commitment protocol, he faces the same opponent in all T periods of the game, so that this protocol makes the game a finitely-repeated PD with matched opponent. If a player is in the no commitment protocol, a new opponent is drawn randomly from the population, and the probability to meet the same opponent again is almost surely zero, making each period a one-shot PD. This means automatically that the strictly dominant strategy for selfish players in the nc condition is "always defect", $s_{nc}^{SE} = \{D, D, \ldots D\}$.

23. Exploring the set of all equilibria, including those employing mixed-strategies, is complex in this game. Allowing for mixed strategies preserves the finding that there are no fully separating equilibria. However, semi-separating equilibria are possible where CC's mix between commitment and no commitment.

B.1.1. Nash Equilibria with Exogenous Commitment

In this section we will establish that in finitely repeated PD's, under both matching protocols, there are both uncooperative and cooperative equilibria, if there are enough CC's in the population. Let us first look at the equilibria in a one-shot PD with random matching. A Nash equilibrium in this game is defined as a pair of actions, $a^* = \left\{a^{SE^*}, a^{CC^*}\right\}$, where a^{SE^*} is the equilibrium action of a selfish player and a^{CC^*} is the equilibrium action of a CC. A selfish player has a strictly dominant action, D, which means that $a^{SE^*} = D$.

For CC's, there are two possible Nash Equilibrium strategies in the stage game. Like the selfish players, CC's may defect, as their best response to defection is also defection. However, their best response to cooperation is also cooperation; so if all CC's cooperate *and* the probability b of being matched with another CC is high enough, then cooperation is an equilibrium strategy for CC's in the stage game. The threshold for the fraction b of CC's in the population is given by $\underline{b} = -q/(1-q-r')$. Below this threshold, the only equilibrium in any one-shot or finitely repeated PD is universal defection by both types. In the following, we will focus our analysis only on the case where $b \geq \underline{b}$. The two possible Nash equilibria in the stage game are then: $\{D, D\}$ and $\{D, C\}$.

We will now define the Nash equilibrium for a T-period PD without commitment. In this case, the Nash equilibrium is a pair of vectors, $s^* = \left\{s^{SE^*}, s^{CC^*}\right\}$, both of length T, and each element of the vectors, $a_t^{SE}(b, h_t^i)$ or $a_t^{CC}(b, h_t^i)$, describes a type's action in the respective period, given b and the personal history of the player. Furthermore, for ease of notation, let $s^*(\tau)$ denote the pair of strategy vectors containing the first τ elements of s^{SE} and s^{CC}, respectively.

As explained above, the nc condition is a string of one-shot PD's, so selfish players will always play the strictly dominant strategy, defect: $s_{nc}^{SE^*} = \{D, D, \ldots D\}$. For CC's, an equilibrium strategy is defined by $s_{nc}^{CC^*} = \left\{a_1^{CC^*}(b), a_2^{CC^*}(b, h_2^i), \ldots, a_T^{CC^*}(b, h_T^i)\right\}$, where $a_t^{CC^*}(\cdot) \in \{D, C\}$. In other words, the CC's may have an equilibrium strategy that prescribes defection in every period, cooperation in every period, or a strategy that contains both defection and cooperation in different periods.

An informative value for the analysis of endogenous commitment will be the *expected*

rate of cooperation, $\hat{c}\left(s^*\right)$:

$$\hat{c}_{nc}\left(s^*_{nc}\right) = b\frac{1}{T}\sum_{t=1}^{T} a_t^{CC*}\left(b, \hat{h}_t\left(s^*_{nc}(t-1)\right)\right),$$

where $\hat{h}_t\left(s^*(t-1)\right)$ is the expected history that is produced by the given equilibrium strategies, s^*_{nc}, up to period $t-1$. Note that, since selfish players will never cooperate in the nc condition, $0 \leq \hat{c}_{nc}\left(s^*_{nc}\right) \leq b$. The expected per-period payoff of a selfish player, who always defects in nc, given the equilibrium s^*_{nc}, is $E\pi^{SE}_{nc}\left(s^*_{nc}\right) = r\hat{c}_{nc}\left(s^*_{nc}\right)$, and $0 \leq E\pi^{SE}_{nc}\left(s^*_{nc}\right) \leq rb$. Similarly, the CCs' expected per-period payoff for any equilibrium strategy is $E\pi^{CC}_{nc}\left(s^*_{nc}\right) = q + (1-q)\hat{c}_{nc}\left(s^*_{nc}\right)$, and this value is bounded by $0 \leq E\pi^{CC}_{nc}\left(s^*_{nc}\right) \leq q + (1-q)b$. Note that $b > \underline{b}$ implies $(1-q)b + q > 0$, meaning the CC's' expected payoff is positive. In contrast, in an equilibrium prescribing universal defection, their expected payoff is zero.

The c condition matches players with a fixed opponent for all T periods. Since in the stage game, defection is the dominant strategy for selfish players and is also an equilibrium strategy for CC's, universal defection is also a (subgame perfect) Nash equilibrium: $s_c^{defect*} = \{\{D, D, \ldots, D\}, \{D, D, \ldots, D\}\}$. As has been shown by Kreps et al. (1982), there also exist equilibria with cooperation as an equilibrium outcome, provided that the probability of the opponent being a CC is sufficiently high. Consider, for example, the following pair of strategies: CC's play "grim trigger" by starting with C in the first period and then continuing with C unless they have observed D at least once in their history; if they do, they switch permanently to playing D. Selfish players have the same strategy, except that, in the last period, they defect with certainty. Kreps et al. (1982) call this behavior "rational cooperation". This pair of strategies is a subgame perfect Nash equilibrium if two conditions are met: first, because CC's are required to cooperate in the final period, even though the selfish players defect, the above condition $b \geq \underline{b}$ has to be satisfied; second, for selfish players it has to be rational to cooperate in all periods but the last if the opponent cooperated as well up to that point. In particular, in the second to last period, $T-1$, a selfish player has to trade off whether he wants

to stick to the equilibrium strategy and earn 1 in $T-1$ and an expected br in period T against defecting in $T-1$, giving him r in that period and 0 in T. Therefore, it is necessary that $1+br \leq r$ or $b \geq 1-1/r$ for this pair of cooperative strategies to be an equilibrium.

The above example is the equilibrium with the highest cooperation rate, as both types cooperate up to the second-to-last period. In the final period, SE's will always defect as the continuation value of their play drops to zero.[24]

B.1.2. Nash Equilibria with Endogenous Commitment

In the previous section, we have established that, with enough CC's in the population, there are both uncooperative and cooperative equilibria in the finitely-repeated PD with exogenous matching protocol. We now turn to the main issue of this paper, the situation where players can *choose* whether they play the PD in the c condition or the nc condition. To investigate this situation, we extend the above model so that players can specify, *ex ante*, which matching protocol they would like to have implemented when playing the finitely-repeated PD. We will represent this decision by a variable that takes on the value 1 if the commitment protocol is chosen and 0 if the no commitment protocol is chosen. Call this binary variable α for SE's and β for CC's. Whenever possible, players update their beliefs about the prevalence of CC's according to these choices and form the two Bayesian posteriors

$$b'_c = \frac{\beta b}{\alpha(1-b)+\beta b} \qquad b'_{nc} = \frac{(1-\beta)b}{(1-\alpha)(1-b)+(1-\beta)b},$$

where b'_c is the probability of being matched with a CC in the c protocol, given b, α and β, and b'_{nc} is the respective probability in nc.

The natural solution concept in this framework is Perfect Bayesian Nash Equilibrium (PBNE). We define a PBNE as a pair of strategy profiles: α^*, s_{nc}^{SE*}, s_c^{SE*} for SE's, and β^*, s_{nc}^{CC*}, s_c^{CC*} for CC's. This pair is a PBNE if the strategies are mutual best responses

24. Allowing for mixed strategies allows for many more equilibria where selfish players start out cooperating and switch to a mixed stage-game strategy at some point.

given the equilibrium beliefs, and the belief system b'_c and b'_{nc} is consistent with Bayes' rule as defined above, whenever it can be applied. The key question in this section is which decisions α and β are part of a PBNE, or in words, which commitment options do the different types choose? Our intuitive claim is that, leaving aside the trivial case where nobody ever cooperates, in equilibrium, both types will choose the same commitment option.

Proposition 1

Any PBNE equilibrium that is cooperative involves pooling.

Proof. We will prove this in two steps. First, we prove by construction that there exist pooling equilibria for both commitment options. Then, we will rule out separating equilibria by contradiction.

(1) Existence of pooling equilibria We first construct a simple cooperative equilibrium for pooling on no commitment, that is, $\alpha = \beta = 0$. Suppose that CC's always defect after choosing commitment, and that they always cooperate in no commitment. The SE's best-respond by always defecting in both commitment options. Both types choose no commitment with certainty. SE's have no incentive to deviate from defection and CC's have no incentive to deviate from cooperation, since $b'_{nc} = b$ (recall that we look only at non-trivial cases where $b \geq \underline{b}$). Likewise, no type has an incentive to deviate to the "commitment" option, regardless of the out-of-equilibrium belief b'_c because the expected payoff there is zero, as both types deviate.[25]

Second, we construct a cooperative equilibrium for pooling on commitment, so $\alpha = \beta = 1$. To this end, we refer back to the equilibrium example we gave for the exogenous commitment case above. Suppose now that CC's always defect in nc but play a (conditionally cooperative) grim trigger strategy in the c protocol. SE's best respond by always defecting in nc and cooperating rationally up to period $T-1$ in c, and then defecting in

25. If nobody chooses to commit, the game is not defined for individual deviation to the commitment option, as there are no opponents. We can accommodate this issue by assuming that there is an infinitesimally small exogenous probability that the commitment choice is reversed. If we further allow this probability to vary by player type, we can generate any Bayesian off-equilibrium belief $0 < b'_c < 1$.

the final period. In the commitment choice stage, both types choose commitment with certainty, hence $b'_c = b$. As we have shown, SE's do not have an incentive to defect before the final period iff $b \geq 1 - 1/r$. Again, independent of the out-of-equilibrium belief, no type has the incentive to deviate to the outside option, nc, as both types defect there.[26]

(2) No separating equilibria Next, we show that there cannot be any cooperative equilibria where one type chooses one commitment option and the other type the other option, so $\beta = 1 - \alpha$. Because of the strict separation, opponents' type is common knowledge among players. That is, all players know that in one option, every player will be of one type and in the other option, every player will be of the other type (that is, $b' = 0$ for the option chosen by the SE's and $b' = 1$ for the option chosen by the CC's). Since SE's only have an incentive to cooperate in the commitment option, and then only if there is a sizable probability of being matched with a CC, this means that SE's will defect with certainty in "their" option, giving them zero payoff. On the other hand, CC's can be certain to be among themselves and are supposed to exhibit some cooperation (because we are looking for a cooperative equilibrium). Furthermore, because this is a Nash Equilibrium, this payoff has to be positive (as switching to always defect yields a minimum payoff of zero). But this provides an incentive for the SE's to join the cooperative CC's: by switching to the other commitment option and defecting, they can secure a strictly positive expected payoff, a contradiction. ■

B.1.3. Irrelevance of Endogenous Choice for PD

Since in pooling equilibria everybody chooses the same commitment option, players' updated belief will just be equal to the proportion of CC's in the overall population, b. But this, together with the requirement of the PBNE that the PD strategy must be a best response given the updated belief, means that the set of equilibria in the PD stage of the game has to be identical to the set of equilibria of the PD with exogenous

26. Recall that the set of players is defined as a continuum. This is convenient as it ensures that there is an infinite number of opponents to draw from in the off-equilibrium nc option, but the number of periods, T, is finite. Thus, the game in the nc condition remains a string of one-shot PD's as long as it is chosen by a positive proportion of players.

commitment.

Corollary 1 (Irrelevance of Endogenous Commitment for PD Behavior) *In all Perfect Bayesian Nash Equilibria of the game with endogenous commitment choice, the set of possible Nash Equilibria in the entailing repeated PD is the same as in the repeated PD with the same exogenously imposed duration.*

Universal defection is an equilibrium under all exogenous and endogenous commitment durations. The two cooperative equilibria in the endogenous game involve pooling, which means $b = b'_c = b'_{nc}$. So, for a fixed b, Nash equilibria of the T-period PD following the commitment choice must be the same as those in the T-periods PD with the same exogenously-imposed interaction duration. Thus, endogenous commitment choice is irrelevant with respect to the equilibria that result the ensuing PD.

B.1.4. Relative Attractiveness of Equilibria

Note that CC's' expected payoff is strictly higher in the pooling equilibrium involving commitment than in the one involving no commitment, independent of the proportion of CC's, b:

$$E\pi_c^{CC^*} = T - 1 + b + (1-b)q \quad > \quad E\pi_{nc}^{CC^*} = T\left(b + (1-b)q\right)$$

Also, the difference between the two expected payoffs grows linearly with the number of periods, T. So, for CC's, the commitment equilibrium is the payoff dominant option, and more so with a growing interaction horizon.[27] Moreover, note that the variable part of the payoff in the commitment equilibrium is the one Bernoulli trial in the final period, where a CC either encounters another cooperating CC and receives payoff 1 (with probability b), or encounters a selfish player and receives q (with probability $1-b$). In contrast, a CC faces this same Bernoulli trial in *all T periods* of the PD in the no-commitment equilibrium. The payoff variance, therefore, is lower in the commitment equilibrium than in the no-commitment equilibrium, and again, this difference grows linearly with the number of

27. Recall that an uncooperative equilibrium yields zero payoff with certainty.

periods, T. Indeed, a CC can only make lower payoffs in the commitment equilibrium than in the no-commitment equilibrium if *all* the opponents in the no-commitment game are CC's, and the opponent in the commitment game is a selfish player. The probability of this event, $(1-b)b^T$, declines exponentially with growing time horizon. In short, for CC's, pooling on commitment is more attractive than pooling on no commitment, and vastly so if T is large.

Selfish players, on the other hand, may prefer the no-commitment equilibrium in terms of expected payoff. This is the case if there are enough CC's in the population so that the probability of earning r in any given period of the no-commitment game outweighs the certain payoff of 1 in any of the non-final periods of the commitment game. Formally, this is the case if $b > 1/r$. These observations regarding the relative attractiveness of equilibria with different commitment options are the basis for our behavioral prediction of initial sorting according to player type.

B.2. Initial sorting with bounded rationality

Following our earlier analysis, we should not expect separation between selfish and conditionally cooperative players in commitment choices. This stands in contrast with the intuition guiding our research—that, given the ability to select different commitment options, CC's will find longer-term commitment more attractive. Therefore, we now show how a simple model of boundedly-rational behavior can provide an—admittedly, *ad hoc*—formal basis for this prediction. One way of viewing this analysis is that the above equilibrium predictions are what we expect in the long-run, but that limited strategic sophistication may yield some separation between types in commitment duration, at least initially. To make the point, we look at the two-period version of the above game and introduce a simple form of bounded rationality.

We adopt a simplified Level-k framework, in which we assume some behavior on the part of unsophisticated (Level-0) players, and then iterate best response over higher levels of sophistication to generate behavioral predictions. Following other research (e.g., Crawford and Iriberri 2007; Crawford, Gneezy, and Rottenstreich 2008), we only make

an assumption regarding the behavior of the Level-0 players.

Specifically, returning to the two player types that we defined earlier—selfish and conditionally cooperative (CC)—assume that selfish types defect in the one-shot game but cooperate rationally under commitment and that CC's cooperate generally but, in the commitment option, switch to defection iff their opponent defects ("grim trigger").[28] We also assume that half of the L0 players select one commitment option, and the other half select the other option, independent of their type. As a consequence, the fraction of CC players in the population of L0 players is equal to their proportion in the overall population, b.

Following the literature, L1 players are then assumed to select a best response to the behavior of L0 players, with knowledge of the proportion of CC types, or b. Similarly, L2 players select a best response to the behavior of L1 players.

Proposition 2 (Existence of Sorting with Bounded Rationality)
Suppose that (i) L0 players select commitment choices non-strategically and are equally likely to opt for either choice, (ii) L0 selfish types play the PD strategy "always defect" in nc and "rational cooperation" in c, and (iii) L0 conditionally cooperative types play the PD strategy "always cooperate" in nc and "grim trigger" in c. Then L1 selfish types will opt for no commitment iff $b > 1/r$, and L1 CC types will opt for commitment. L2 and higher players of both types will pool on commitment.

Proof. This proposition uses the behavioral concept of finite levels of reasoning and iterated best response. That is, players play optimally in the PD given their social type and their belief, but they do not form this belief fully rationally. Instead, a L1 player best responds to the belief that players choose the two commitment options equally often and non-strategically, i.e., independent of type. This means that a L1 player expects to encounter CC's in *both* options with probability b. A L1 player also expects CC's to

28. Note that, when b is high enough to support cooperation, the PD behavior yields weakly higher payoffs for both players than the other PD equilibrium strategies, making it a natural choice for unsophisticated players. This case reflects the more general requirement for our behavioral prediction to hold and on which our intuition is based, that strategically unsophisticated (L1) selfish types believe that there are (L0) CC's who choose the no commitment option but nevertheless cooperate.

play the PD strategies "always cooperate" in the nc protocol and "grim trigger" in the c protocol, and selfish players to play "always defect" in the nc protocol and "rational cooperation" in the c protocol.

Anticipating these L0 strategies, an L1 CC player will strictly prefer the commitment option and the c protocol strategy "grim trigger" because she can secure a payoff of 1 in the first period and prefers, in expectation, cooperation over defection in the second period as long as $b > {}^{-q}/_{1-q-r'}$. A selfish L1, however, will prefer the no commitment option iff there are enough CC's in the population, to yield a higher expected payoff in the first period of the one-shot option (br) than in the commitment option (1). So if the selfish L1 players have an "optimistic" prospect of encountering many CC's in either option (i.e., $b > {}^{1}/_{r}$), they will prefer no commitment and will defect in both periods of the PD. Thus, the two L1 types separate into two different commitment options.

L2 players best-respond to the choices of L1 players. Because, for $b > {}^{1}/_{r}$, selfish L1 types choose the one-shot option and defection, but L1 CC's choose to commit and play "grim trigger", L2 players of either social type expect to encounter only defecting selfish types if they choose no commitment; by the same argument, they expect to find CC's playing a trigger strategy if they choose to commit. This means that L2 CC's will also choose to commit. Moreover, *selfish* L2 players will also choose to commit and select rational cooperation (C^R) in the second stage. ■

C. Subject Instructions

We present the complete instructions for the CHOICE-NIP condition, because this was the most complex one. The instructions for EXO-N, EXO-I, EXO-P, CHOICE-NI, comprehension questions, questionnaire, and ztree files are available upon request from the authors.

Initial Instructions

Thank you for participating in today's experiment.

I will read through a script to explain to you the nature of today's experiment as well as how to navigate the computer interface with which you will be working. I will use this script to make sure that the information given in all sessions of this experiment is the same.

In addition to a 10 CHF payment that you receive for your participation, you will be paid an amount of money that you accumulate from the decision task that will be described to you in a moment. The exact amount you receive will be determined during the experiment and will depend on your decisions and the decisions of others. You will be paid privately, in cash, at the conclusion of the experiment.

All monetary amounts you will see in this experiment will be denominated in ECUs or Experimental Currency Units. We will convert ECUs into CHF at the rate of

150 ECUs = 1 CHF.

If you have any questions during the experiment, please raise your hand and wait for an experimenter to come to you.

Please do not talk, exclaim, or try to communicate with other participants during the experiment.

Do not use the computer in a way not specified by these instructions or by the experimenters.

Participants intentionally violating the rules may be asked to leave the experiment with only their participation payment.

The Task

In each stage of this experiment, you will participate in a simple decision task. Stage 1 will consist of 1 period of this task, while Stage 2 will consist of 150 periods of this task. What happens in each stage will not affect the procedures or your earnings in the other stage.

In every period of this experiment you will be paired with one other participant. You will be paired with this participant through a computer network. At no time will your true identity be revealed to the other participants, nor will you ever know the identity of the participant with whom you are paired. How you are paired with the other participant will be explained later.

Both you and the other participant will have two possible choices. You can choose A or you can choose B.

- If you **both choose A** you will both receive payoffs of 40 ECUs.
- If you **both choose B** you will both receive payoffs of 25 ECUs.
- If **you choose B but the other participant chooses A**, you will receive a payoff of 65 ECUs and the other participant will receive 5 ECUs.
- If **you choose A but the other participant chooses B**, you will receive a payoff of 5 ECUs and the other participant will receive 65 ECUs.

These payoffs are also summarized in the table below. The **bold** number in the bottom-left portion of each box is the payment received by **you**, the number *in italics* in the top-right portion is the payment received by the *other participant*:

		Other's choice:	
		A	*B*
Your choice:	**A**	*40* **40**	*65* **5**
	B	*5* **65**	*25* **25**

In every period of the experiment, you and the other participant with whom you are paired will select between a choice of A or B for that period. Once all participants in the experiment have made their choices, your payoff will be determined by the combined choices of you and the person with whom you are paired.

Stage 1

In Stage 1, you will participate in the decision task that we just described to you *for 1 period*. For this purpose, you will be randomly paired with another participant in the room.

In Stage 1, you will make your decisions *sequentially*. That is, one person in each pair will decide first whether he or she chooses A or B. The participant moving first does not know which decision the second participant will make. The other participant will then make his or her decision of A or B **knowing** what the first participant has chosen.

The computer will select **at the end** of the experiment who decides first and who decides second. This means that when you make your decision in Stage 1, you do not yet know whether you are the one deciding first or the one deciding second. Therefore, in the 1 period of Stage 1, you will indicate three choices:

- First, you will indicate whether you choose A or B **in case you decide first**
- Then, you will indicate whether you choose A or B **in case you decide second after the other participant has chosen A**
- Then, you will indicate whether you choose A or B **in case you decide second after the other participant has chosen B**

At the end of the experiment, the computer will randomly select either you or the other participant with whom you are paired to be the first to decide. It will then take this person's first decision and match it with the corresponding decision of the other participant, who decides second.

For example, if the computer decides that you are the one who decides first, and you chose A in that case, the computer will then look up what the other participant chose in case he/she decides second after you have chosen A. The computer will then compute your respective payoffs according to these two choices. At the end of the experiment, you will be informed about who decided first, the choice made by the person with whom you are paired, and the respective outcome.

Stage 2

Stage 2 will consist of 150 periods in which you will participate in the same decision task as before. However, in Stage 2, all participants will make their choices *at the same time*, meaning that no one will know what the person with whom they are paired has chosen when choosing between A and B.

How participants are paired in Stage 2

You will face the one-period task just described for a total of 150 periods, which will be divided into 15 sets of 10 periods each. That is, each set will consist of 10 periods of the task and there will be 15 sets.

In Stage 2, you are in a matching group with 15 other participants. This matching group, consisting of 16 participants, will be fixed for all of Stage 2, and you will only be paired with others in your same matching group. You will never know the identity of anyone in your matching group.

To begin a set of periods, you will first indicate whether

(1) you want to be paired with the *same other participant for the entire remaining experiment*,
(2) you want to be paired *with the same other participant for the 10 periods* in that set, or whether
(3) you want to be newly paired *with a randomly selected participant at the beginning of every period* in the set.

You do this by clicking one of three radio buttons.

- If you select the radio button "one random participant for all remaining sets of the experiment", this means that you want the computer to randomly pair you with another participant at the beginning of the set, and to keep you paired **with this same participant for the rest of the experiment**. That is, all remaining periods in the remainder of the experiment – in this set and in all remaining sets – will be with the same participant.

- If you select the radio button "one random participant for the 10 periods in the set", this means that you want the computer to randomly pair you with another participant at the beginning of the set, and to keep you paired **with this same participant for all 10 periods in the set**. That is, all 10 periods in the set will be with the same participant.

- If you select the radio button, "one random participant for each period in the set", this means that you want the computer to randomly pair you with another participant at the beginning of **every period**. That is, in each of the 10 periods in the set you will be re-paired with another participant.

After all participants indicate how they would like to be paired in that set, the computer will form three lists: one list consisting of all the participants who would like to be paired with the same participant for the entire remaining experiment, one list consisting of all the participants who would like to be paired with the same participant for the 10 periods of the next set, and a third list consisting of all the participants who have chosen to be re-paired at the beginning of each of the 10 periods. If there is an even number of participants in each list, then the computer will pair everyone at random with someone else from the same list.

If there is an odd number of participants who would like to be paired with the same participant for all remaining periods of the experiment, then not all of these choices can be implemented. The computer will then randomly pick one of these participants and move him or her to the list consisting of those who will be paired with the same participant for the next 10 periods. This makes the number of participants in the first list even.

If there is now an odd number in the list of participants who want to be paired with the same participant for the next 10 periods, the computer will randomly pick one of the participants who chose that option and move him or her to the list consisting of those who will be randomly re-paired at the beginning of each of the 10 periods. This makes the number of participants in *all* lists even, and the computer will then randomly pair everyone with someone from the same list.

You will then be informed of whether you will be paired in the way that you selected for that 10-period set. Note that most of you will be paired in the way you selected, but in some cases we may need to change the way you are paired from what you selected, if uneven numbers of people chose each of the three ways.

The computer will then randomly pair each of you with one other participant in your matching group from the list of people who chose the same way in which to be paired for that set.

For those participants who are paired with the same participant for all remaining periods, they will remain paired with the same other participant for the entire remaining duration of the experiment. This also means that they will not get to choose how to be paired at the beginning of future sets. For those participants who are paired with the same participant for all 10 periods in that particular set, they will remain paired

with the same other participant for the entire set. For those participants who are re-paired at the beginning of each period, they will be randomly re-paired with one of the other participants from the same list in each of the 10 periods.

After the 10th period a new set will begin. All participants who have not previously chosen to remain paired with the same participant for all remaining periods will then again have a choice to either be paired with the same randomly selected participant for the entire remaining experiment, be paired with the same randomly selected participant for all 10 periods in the next set, or to be re-paired with another participant in each of the 10 periods in the next set.

At the end of each period, the computer will tell you your choice in that period, the other participant's choice, and your earnings from that period. You will also be able to look over the history of outcomes in prior periods.

Then another period begins. If the period is part of the same 10-period set as the previous period, you will be paired with the same other participant as before or with a new other participant, depending on the choice that you made at the beginning of that set. If the period is the first period of a new set, and you have not previously chosen to remain paired with the same other participant for the rest of the experiment, you will again begin by indicating how you would like to be paired for the 10 periods in that set.

Summary:

- Stage 2 consists of 150 periods, divided into 15 sets of 10 periods
- In each period you will make a choice of A or B at the same time as the person with whom you are paired
- At the beginning of the first set, you will be newly paired with a randomly selected participant; you will decide whether you want to be paired with this same participant for all remaining periods of the experiment, to be paired with this same participant for all 10 periods in that particular set, or to be newly paired with randomly selected participants in every period of that set.
- In case you choose to be paired with the same other participant for the rest of the experiment, you will not get to choose how to be paired at the beginning of future sets. Otherwise, you will face the same decision at the beginning of every set of whether you want to be paired with the same participants for all remaining periods of the experiment, to be paired with the same participant for all 10 periods in that particular set, or to be newly paired with randomly selected participants in every period of that set.